Technologiegestützte Leistungsdiagnostik in Schule und Hochschule

Inaugural-Dissertation
zur
Erlangung des Doktorgrades
der Sozial- und Verhaltenswissenschaftlichen
Fakultät der
Friedrich-Schiller-Universität Jena

vorgelegt von

Nadine Schlomske
aus Freiburg im Breisgau

Wintersemester 2012/ 2013

Erstgutachter	Prof. Dr. Michaela Gläser-Zikuda
Zweitgutachter	PD Dr. Dirk Ifenthaler
Dekan der Sozial- und Verhaltens-wissenschaftlichen Fakultät	Prof. Dr. Stephan Lessenich
Datum des Promotionsbeschlusses	24. Oktober 2012
Datum der Disputation	24. April 2013

Für meine Eltern

Zita & Christian

Danksagung

An dieser Stelle danke ich meiner Doktormutter Frau Prof. Dr. Gläser-Zikuda für die wertvolle Unterstützung und die anregenden Gespräche. Ihre Impulse, was die methodologisch gemischte Herangehensweise anbelangt, haben sehr zum Gelingen dieser Arbeit beigetragen. Ebenso danke ich meinen Zweitgutachter, PD Dr. Dirk Ifenthaler für die kollegiale Kooperation. Die inspirierende Zusammenarbeit mit PD Dr. Pablo Pirnay-Dummer hat die Arbeit in vieler Hinsicht bereichert und immer wieder neu ins Rollen gebracht.

Meinen Kollegen sowie allen studentischen Hilfskräften am Lehrstuhl für Schulpädagogik und Didaktik an der FSU Jena danke ich für den wertvollen Austausch. Dabei geht ein ganz besonderer Dank an Julia Rohde, Michael Wiegleb (M.A.), Dr. Sascha Ziegelbauer, Susi Limprecht (M.A), Jan Fendler, Dr. Gloria Hempel, Christina von Obstfelder (M.A.), Sebastian Kretzschmar und Vicky Gebhard.

Meiner Chefin Frau Prof. Dr. Seidel danke ich für das große Entgegenkommen, diese Arbeit zu einem guten Ende zu führen.

Ein weiteres Dankeschön geht an die Teilnehmer des ASQ-Workshops: Sina Willer, Jessica Schoder, Christin Steinborn, Peggy Seidemann, Sophia Hackethal und Dana Schütze für die wertvolle Zusammenarbeit bei den Auswertungen der Interviews

Ich bedanke mich bei allen Schulen, die an dem Projekt teilgenommen haben. Insbesondere bei den Schülern, Eltern, Lehrern und Direktoren sowie allen mitwirkenden Studenten der Friedrich-Schiller-Universität Jena.

Kristin Stepper, Nadia Bickel, Katrin Fischer und Stephan Preißler danke ich für das freundschaftliche Mittragen der Arbeit.

Von Herzen danke ich meiner Familie, für die kontinuierliche Ermutigung und meinem Freund Ralf Bodenstein für die liebevolle Unterstützung beim Abschließen dieser Arbeit.

In der vorliegenden Arbeit wird aus Gründen der Lesbarkeit auf die reihende Listung von fast identischen Substantiven verzichtet, wenn Gattungsbegriffe verwendet werden, denen keine geschlechtsspezifische Bedeutung (z. B. Proband) zugedacht ist. Wenn geschlechtsspezifische Aussagen gemacht werden, so werden die entsprechenden Substantivformen (z. B. Schülerin und Schüler) verwendet. Alle wissenschaftlichen Notationen sind in Anlehnung an das APA Publication Manual realisiert (American Psychological Association, 2007).

INHALTSVERZEICHNIS

INHALTSVERZEICHNIS

1 Einleitung

Leistungsmessung und Bewertung spielen im Schul- und Hochschulkontext eine zentrale Rolle, was in der selektiven Funktion von Bildungseinrichtungen begründet liegt (vgl. Fend, 2008). Für Lehrkräfte bedeutet das, externalisiertes Wissen von Lernenden auf Grundlage diagnostischer Expertise hinreichend objektiv, zuverlässig, sowie gültig einzuschätzen (vgl. Schrader, 2006). Insbesondere textproduzierende Schülerleistungen (z. B. in Form von Aufsätzen) führen oft zu geringen Bewertungsübereinstimmungen (vgl. Ingenkamp & Lissmann, 2008; Klauer, 1982; Linn, Klein & Hart, 1970).

Die vorliegende empirische Forschungsarbeit ist im Bereich der technologiegestützten Leistungsdiagnose anzusiedeln. Dabei interessiert vor allem die Fragestellung, inwiefern sich die Bewertung von textbasierten Leistungen mithilfe eines innovativen Instruments, welches im Kontext der Forschung zu mentalen Modellen entwickelt wurde, optimieren lässt (Seel, 1991). Für diesen Zweck untergliedert sich die vorliegende Arbeit in zwei wesentliche Bestandteile. Zunächst in die theoretischen Vorüberlegungen sowie die Einbettung in den empirischen Stand der Forschung und schließlich in die empirische Überprüfung im Hochschul- sowie im Schulkontext. Im Nachhinein erfolgt eine Diskussion und Einbettung der Ergebnisse in den aktuellen Forschungsstand.

1.1 Zielsetzung

Die Zielsetzung der vorliegenden Arbeit ist die empirische Überprüfung technologiegestützter Leistungsdiagnostik textbasierter Lernergebnisse. Aus methodologischer Perspektive steht die Frage nach einer zeitökonomischen, objektiven, zuverlässigen und gültigen Bewertung textbasierter Lernergebnisse im Zentrum. Es wird der Frage nachgegangen, ob sich im Rahmen der Theorie der mentalen Modelle entwickelte Verfahren eignen, um die Leistungsfeststellung durch Lehrende (in Form von Noten und Leistungspunkten) zu unterstützen.

1.2 Gliederung der Forschungsarbeit

Das vorliegende Kapitel zeigt die Gliederung der Forschungsarbeit auf. Die Einleitung führt zum Thema hin und beschreibt den Fokus der methodologischen

Überprüfung technologiegestützter Leistungsfeststellung in Bezug auf Zeitökonomie sowie der Berücksichtigung der wissenschaftlichen Standards (Objektivität, Reliabilität und Validität - siehe Zielsetzung). Das z*weite Kapitel* bettet die Zielsetzung in den kognitivistischen Lernbegriff der mentalen Modellbildung und des selbstregulierten Lernens ein. Ein für die Leistungsermittlung wesentlicher Bestandteil ist die vorangegangene Wissensrepräsentation durch den Lernenden. Schließlich folgt die theoretische Verortung in die Expertiseforschung, welche den Rahmen setzt um Lernergebnisse mittels geeigneter Vergleichsmaßstäbe überhaupt überprüfbar zu machen. Im Fokus der Forschungsarbeit steht die diagnostische Expertise von Lehrkräften. Sie ist ein wesentlicher Bestandteil der Lehrerexpertise, um externalisiertes Wissen hinreichend objektiv, reliabel und valide festzustellen. Die diagnostische Expertise wird in den empirischen Forschungsstand eingebettet.

Das *dritte Kapitel* beschreibt die Leistungsdiagnostik durch Lehrende sowie die technologiegestützte Leistungsdiagnostik und diskutiert diese auf Grundlage des aktuellen Forschungsstandes. Das *vierte Kapitel* stellt die in diesen Untersuchungen herangezogenen Instrumente dar. Hier folgt eine methodologisch-kritische Diskussion in Hinsicht auf die Leistungsfeststellung. Das *fünfte Kapitel* zeigt die sich daraus ergebenen Fragestellungen und Hypothesen, welches die Intervention umfasst. Im *sechsten Kapitel* folgt die Methodenbeschreibung zur Überprüfung der gestellten Fragestellungen zunächst im Hochschulkontext und daran anschließend im Schulkontext. Die Ergebnisdarstellung und Diskussion findet getrennt nach Hochschulkontext bzw. Schulkontext statt. *Kapitel sieben* und *acht* erläutern die empirischen Befunde, welche im *neunten* sowie im *zehnten Kapitel* nach Hochschulstudien und Schulstudien getrennt voneinander diskutiert. Das *elfte Kapitel* fasst die zentralen Befunde zusammen und zieht Konsequenzen in Bezug auf eine unterstützende Leistungsdiagnostik.

2 Theoretische Grundlagen

Die Arbeit wird zunächst in den kognitivistischen Lernbegriff der mentalen Modellbildung und des selbstregulierten Lernens eingebettet. Daran schließt sich die Frage der Wissensrepräsentation in Bezug auf die Leistungsmessung an. Danach folgt eine Verortung in den Begriff der Expertise, welche insbesondere die diagnostische Expertise bei Lehrkräften ins Visier nimmt. Sie stellt die Basis für die daran anschließende Darstellung der Leistungsdiagnostik dar, welche in *Kapitel drei* umfassend diskutiert wird.

2.1 Mentale Modelle und Wissensrepräsentation

2.1.1 Mentale Modelle

Das folgende Kapitel beschreibt Lernen auf Grundlage der mentalen Modellbildung. Nach van der Meer (1996) werden Zustände mentaler Repräsentationen an die der Außenwelt angepasst. Dies setzt voraus, dass das kognitive System über die Fähigkeit verfügt, Wissenszustände an die aktuelle Welt anzupassen, damit Lernen stattfinden kann.

Jede Erkenntnis - also Lernen - entsteht durch eine Form von Modellbildung (vgl. Stachowiak, 1973). Diesen Prozess des Erkenntnisgewinns nennt Seel (1991) „mentale Modellbildung“. Mentale Modelle sind kognitive Konstruktionen, die ad hoc und auf Grundlage des semantischen Wissens gebildet werden, um einen bestimmten Weltausschnitt plausibel zu machen (vgl. Ifenthaler, 2006, S. 7). Sie werden nach Johnson-Laird (1983) und Seel (1991) vom Lernenden beibehalten solange sie auf Grundlage des Wissens Plausibilität erzeugen. Dieses konstruierte Modell setzt sich aus jenen Attributen des semantischen Wissens (Weltwissens) zusammen, welche zur Klärung des Weltausschnitts relevant erscheinen (vgl. Stachowiak, 1973, S. 132). Das Weltwissen, über das eine Person in einem bestimmten Gegenstandsbereich verfügt, ist eine Menge an „Sätzen“, die in der Realität entweder zutreffen oder nicht, solange sie vor dem Hintergrund des abgespeicherten Wissens Plausibilität erzeugen (vgl. Seel, 1991, S. 12). Abgespeichertes Wissen kann entweder erfahrungsbegründet oder rational sein. Rationales Wissen schließt Sinneserleben aus (ebd., S. 10). Er differenziert zwischen der erfahrbaren Welt (objektive Realität) und der subjektiven Realität

(Weltwissen) und führt zu der Überlegung, dass ein mentales Modell aus wissenschaftlicher Sicht durchaus falsch sein kann, auch wenn es vor dem Hintergrund des Weltwissens subjektive Plausibilität erzeugt (ebd., S. 49). Das Weltwissen stellt dabei immer eine Abbildung und nie das Original der Welt dar. Nach an der Heiden (1985) setzt das Wissen folgende drei Axiome voraus: 1) die Welt, 2) die „kognitiven Phänomenen“ wie beispielsweise der Wahrnehmung und den Gefühlen und 3) die Unterscheidung zwischen den Inhalten der kognitiven Phänomene und deren Realität in der Welt (vgl. Seel, 1991, S. 10). Diese drei Axiome verdeutlichen, warum die objektive Welt aufgrund der subjektiven Wahrnehmung nie deckungsgleich auf das Weltwissen abgebildet werden kann.

Mentale Modele werden konstruiert, wenn die Welt nicht mehr auf Grundlage eines existierenden Schemas erklärt werden kann. Schemata sind wie mentale Modelle - mit der Ausnahme, dass sie stabil und wieder aufrufbar sind. Schemata bestehen aus bereits gefestigten und somit stabilen Wissensstrukturen, die bereits abgespeicherte und verinnerlichte Handlungsabläufe - beispielsweise einen Restaurantbesuch - beinhalten. „Bleibt die Welt im Rahmen dessen, was bereits erfahren wurde und verhält sie sich den Erwartungen entsprechend, so können gespeicherte Schemata zur Repräsentation herangezogen werden“ (Pirnay-Dummer, 2006, S. 6). Sie werden im kognitiven System assimilativ durch Hinzufügen neuer Leerstellen verändert. Leerstellen verweisen auf Variablen, die im Schema enthalten sind. Bei einem Restaurantbesuch verweisen sie auf Handlungsabläufe, die darin mit enthalten sind, zum Beispiel bei der Bestellung eines Menüs. Ein Assimilationsprozess setzt dabei ein bereits vorhandenes Schema voraus (vgl. Seel, 1991, S.45). Rumelhart & Norman (1978) sind dabei typische Vertreter des schematheoretischen Ansatzes und der Schemarepräsentation. Mentale Modelle können über die Zeit zu Schemata werden, wenn sie immer wieder auf ähnliche Weise konstruiert werden (vgl. Hanke, 2006).

Nach Piaget (1976) verändern Lernende ihr Wissen, indem sie neue Wissensinhalte in ihr bereits vorhandenes Wissen integrieren. Dies führt zu assimilativen Prozessen, solange die neuen Wissensbestandteile nicht im Gegensatz zu dem bereits vorhandenen stehen. Lässt sich ein bestimmter Weltausschnitt vom kognitiven System nicht in bestehendes Wissen integrieren (assimilieren), wird durch mentale Modellbildung eine Wissensveränderung ermöglicht (vgl. Seel, 1991, S. 44). Die Bildung mentaler Modelle führt dazu, dass das

kognitive System „neues“ Wissen in bereits vorhandenes Weltwissen integriert. Es führt also zu akkomodativen Prozessen, in denen bereits vorhandenes Wissen verändert wird. Dies ist der Bereich, in dem Lernen stattfindet.

2.1.2 Selbstreguliertes Lernen

Welchen Einfluss Lernende auf die Lernsituation sowie die abschließende Leistungsfeststellung nehmen können, verdeutlicht der theoretische Rahmen des selbstregulierten Lernens.

Schmitz (2001) entwickelte beispielsweise das Modell des selbstregulierten Lernens, welches auf dem Ansatz von Zimmermann (2000) beruht und von Schmitz & Wiese (2006) weiterentwickelt wurde (vgl. Schmitz & Schmidt, 2007, S. 10). Der Lernprozess gliedert sich nach diesem Modell in drei wesentliche Bereiche. Zunächst die Planungsphase, in der Lernprozesse vorbereitet und Ressourcen überprüft werden. Anschließend findet die eigentliche Lernphase statt, in der Lernstrategien eingesetzt werden. Gläser-Zikuda (2007) weist auf die Notwendigkeit von Lernstrategien und Metakognition in Bezug auf die erfolgreiche Veränderung von Wissen hin (ebd., S. 113). 2006). Typische Vertreter der Lernstrategien sind beispielsweise Friedrich & Mandl (2006), die einen Überblick an empirischen Studien im Bereich der Lernstrategien geben. Lernende, die Mängel in Bezug auf ihre Lernstrategien oder metakognitive Fähigkeiten haben, können in der aktionalen Phase des selbstregulierten Lernens nicht erfolgreich Einfluss auf die Wissensveränderung vornehmen. Zum Schluss folgt die Bewertungsphase, in der der Lernende seine eigenen Lernprozesse reflektiert und seine zuvor gesetzten Ziele (Soll-Werte) mit den erreichten Ergebnissen (Ist-Werte) vergleicht. Liegen Diskrepanzen vor, so kommt es zu einer erneuten Planung des weiteren Lernverlaufs. Dies impliziert, dass Lernende ihren Lernprozess aktiv beeinflussen. Im Idealfall werden sie, wenn es um die Feststellung von Lernergebnissen geht, mit eingebunden. Insbesondere im Fall der Leistungsbeurteilung würde dies beispielsweise das Erstellen der Kriterien mit einschließen. Pirnay-Dummer (2012) verweist an dieser Stelle auf den Einfluss, den die Akzeptanz von Bewertungskriterien bei Lernenden hat. Werden diese von den Lernenden nicht verstanden und nicht akzeptiert, so wirke sich dies negativ auf die Regulation aus (ebd., S. 76). Straka (2006) unterscheidet bei den Lernstrategien

im Kontext des selbstregulierten Lernens zwischen folgenden Ansätzen: 1) die Analyse von Lehr-Lern-Prozessen (vgl. Weinstein & Mayer (1986), 2) die Dimension der Motivation (vgl. Pintrich, 1988), 3) das Drei-Schichten-Modell (nach Boekaerts, 1999), 4) die sozial-kognitive Perspektive (vgl. Zimmermann, 2005) und 5) das mehrdimensionale Strukturmodell. In der vorliegenden Forschungsarbeit finden die motivationalen Voraussetzungen der Lernenden keine Beachtung, da die Analyse bereits realisierter Abbildungen von Wissen im Vordergrund steht. Demzufolge finden das zweite und das fünfte Modell keine Berücksichtigung.

2.1.3 Wissensrepräsentation

Dieses Kapitel beschäftigt sich mit der Repräsentation von Wissen. Dabei interessiert die Frage, in welchen Formaten das kognitive System Wissen mental repräsentiert und in welchem Format es dieses abbildet. Denn das Format, in welchem das kognitive System Wissen abbilden kann, bestimmt das Instrument, welches zur Analyse herangezogen werden kann. Bislang blieben jedoch unerforscht, wie mental abgespeicherte Modelle in eine Repräsentation umkodiert werden.

Erst wenn Lernende ihr Wissen abgebildet haben, können Lernprozesse sichtbar gemacht werden „Über die genaue Form der Wissensrepräsentation im Gedächtnis sowie über Prozesse der Aufnahme und des Abrufs von Wissen bestehen […] eine ganze Reihe verschiedener Modellvorstellungen (vgl. Beckenkamp, 1995; Engelkamp, 1994; Schnotz, 1994; Wessels, 1994)“ (Eckert, 1998, S.9). Bruner (1964) und Aebli (1981) unterscheiden zwischen *enaktiven*, *ikonischen* und *symbolischen* Wissensrepräsentationsformaten (vgl. Ifenthaler, 2006, S. 24). Die enaktive Repräsentation bildet Handlungsabläufe ab, welche ein Ziel beinhalten. Zum Beispiel: „man weiß, was man tun muss, um eine Kurznachricht zu versenden“. Die ikonische Repräsentation stellt Handlungsabläufe in Form von Bildern dar, vergleichbar mit einer skizzenhaften Wegbeschreibung, um von A nach B zu gelangen. Die symbolische Repräsentation bildet Zeichen ab, welche zur Repräsentation des Wissens notwendig sind, zum Beispiel das Alphabet oder Formeln. Die Welt, wie sie von einer Person wahrgenommen und verstanden wird, wird im Gedächtnis abgebildet. Dieser Vorgang lässt sich mit der Abbildfunktion

von Seel (1991) beschreiben. Es erfolgt eine Transformation der objektiv (erfahrbaren) Welt auf das „subjektiv“ wahrgenommene Wissen. Dies wird durch gewisse Zuordnungsvorschriften realisiert.

> „In den Zuordnungsvorschriften D (f) wird durch f die Welt auf ‚Wissen‘ abgebildet. Das Ergebnis wird in den Vorschriften B (f) = D (g) realisiert und wird als ‚Wissen‘ bezeichnet, das durch die Abbildung g in B (g) als ‚Repräsentation‘ organisiert ist“ (Pirnay-Dummer, 2006, S. 11).

Nach Seel (1991) erlangt Wissen erst durch Externalisierung einen symbolischen Charakter (S. 16), da es zur externen Abbildung von intern abgebildetem Wissen symbolischer Zeichensysteme bedarf. In der vorliegenden Arbeit werden dabei die bereits realisierten Abbildungen von Weltwissen untersucht.

Nach empirischen Befunden von Couné, Hanke, Ifenthaler & Seel (2003, 2004) zufolge haben Lernstrategien und allgemein kognitive Fähigkeiten keinen Effekt auf die Qualität der Modellbildung (vgl. Ifenthaler, 2006, S. 78). Die zugrunde liegenden Lernstrategien und motivationalen Bedingungen von Lernenden bei der Externalisierung ihres Weltwissens finden in der vorliegenden Untersuchung keine Berücksichtigung.

Um Lernende dazu zu bringen, ihr Wissen durch symbolische Zeichen sichtbar zu machen, bedarf es einer zuvor gestellten Aufgabenstellung. Lienert & Raatz (1994) geben einen Überblick über die klassischen Aufgabenformate. Dabei unterscheiden sich diese in offene und geschlossene Antwortmöglichkeiten (vgl. Klauer, 2002, S. 105). In der textgenerierenden Aufgabenstellung externalisiert der Lernende sein Wissen in dem er über die Sprache eigene Formulierungen wählt (ebd., S. 105). Lienert & Raatz (1994) betonen, dass textproduzierende Aufgabenstellungen (in Form von Kurzaufsätzen) dem Bewerter es deutlich erschweren diese eindeutig zu analysieren, da eine Vielzahl an Bewertungskriterien mit berücksichtigt werden müssen (ebd., S. 28). Demzufolge ergibt sich möglicherweise eine anspruchsvollere Festlegung einer eindeutigen Musterlösung. Dies liegt auch darin begründet, dass eine relativ offen gehaltene Aufgabenstellung, die mehrere Lösungsvorschläge ermöglicht, eine einzige Musterlösung ausschließt.

Vor dem kognitions- und lernpsychologischem Hintergrund wird zwischen deklarativem sowie prozeduralem Wissen unterschieden. Während das deklarative Wissen die Wissensinhalte fokussiert, beinhaltet das prozedurale Wissen die

Performanzebene (vgl. Seel, 2003). Beobachtet werden kann dabei immer nur das externalisierte Wissen.

Helmke & Weinert (1997) unterscheiden zwischen dem deklarativen, dem prozeduralen sowie dem metakognitiven Wissen. Ersteres fokussiert das in Form von Sprache externalisierbare Wissen. Das prozedurale Wissen fokussiert die zur Lösung einer Aufgabenstellung benötigten Strategien. Letzteres beinhaltet das Wissen über das Lernen, den Abruf sowie der Nutzung des internalisierten Lerninhaltes (ebd., S. 75). Die vorliegende Arbeit fokussiert die ersten beiden Bestandteile.

Pirnay-Dummer (2012) zieht die Konsequenz, dass Lernen aus kognitiver Perspektive immer eine Veränderung hinsichtlich der kognitiven Strukturen ist (ebd., S.32). Die Lehr-Lern-Forschung beschäftigt sich mit dem Aufbau und der Reorganisation von Wissensstrukturen und „mit der didaktisch optimal aufbereiteten Vermittlung von Fertigkeiten und den komplexen Zusammenhängen zwischen Lehren und Lernen“ (Ifenthaler, 2006, S.57). Das Lernen, welches beispielsweise in der Schule, an der Universität oder in der Berufsausbildung stattfindet, hat dabei im vorliegenden Kontext die Zielsetzung, „dass eine Person nach dem Lernprozess in einem bestimmten Bereich höhere Kompetenz besitzt als vorher“ (Gruber & Mandl, 1996, S. 583). Nach Willett (1989) umfasst Lernen *Wachstum* und *Veränderung*, was insbesondere in Schulen gemessen werden sollte ebd., S. 346). Lernumgebungen (Lehre) haben das Ziel Lernprozesse zu initiieren, zu fördern und zu begleiten (vgl. Seel, 2010, S. 5). Lernprozesse zu initiieren hat demzufolge das Ziel, den Lernenden auf einen höheren Expertisegrad zu bringen. Das Konstrukt der Expertise wird in 2.1.4 dargestellt und deren Relevanz für die vorliegende Arbeit verdeutlicht.

Das folgende Kapitel beschäftigt sich mit den empirischen Befunden der Expertiseforschung und fokussiert die diagnostische Expertise, welche Lehrkräfte für die Analyse von Lernergebnissen benötigt.

2.2 Expertiseforschung und Lehrerexpertise

2.2.1 Novizen- und Expertiseforschung

In der Expertiseforschung wurden Novizen bislang gerne als Vergleichsgruppe zu Experten herangezogen, da sie einen geeigneten Kontrast bilden (vgl. Gruber 1994, S. 17). Experten unterscheiden sich von Novizen nicht nur in der Menge des verfügbaren Wissens, sondern auch in deren Strukturierung und Verknüpfungsdichte (vgl. Pirnay-Dummer, 2006, S. 5). Im Gegensatz zum Novizen erzielt ein Experte nach Posner (1988) dauerhaft hervorragende Leistung (ebd., S. 584). Diese Vorteile kann er jedoch nur in der jeweiligen Domäne, in der er Experte ist, vorweisen (vgl. Gruber & Mandl, 1996, S. 585). Gruber (1994) unterscheidet zwischen vier möglichen Bereichen, in denen sich bei Experten hohe Leistungen feststellen lassen: 1) manuelle Tätigkeiten, 2) mentale, akademische Tätigkeiten, 3) komplexe Tätigkeiten und 4) künstlerische Tätigkeiten (vgl. Gruber, 1994, S. 10 f). In der vorliegenden Arbeit sind vor allem der zweite und der dritte Bereich von Interesse. Nach Posner (1988) hat der Novize - im Vergleich zu Experten - Leistung entweder noch nicht erreicht oder kann diese nicht erreichen (vgl. Gruber & Mandl, 1996, S. 584 f). Möglicherweise fehlen ihm dazu die notwendigen kognitiven Strategien. Ferner sind Novizen laut Gruber (1994) dadurch gekennzeichnet, dass sie auf einem bestimmten Gebiet noch keine Erfahrungen gesammelt haben (ebd., S. 10). Experten verfügen in einem bestimmten Gegenstandsbereich über „wissenschaftliche“ Modelle, Laien hingegen greifen auf „Alltags-Modelle“ zurück (vgl. Seel, 1991, S. 7). Nach Pirnay-Dummer (2006) lassen Expertenmodelle „die meisten validen Schlussfolgerungen, über die Welt in der spezifischen Domäne, des Gegenstandsbereichs zu einer gegebenen Zeit“ zu (ebd., S. 13). Um den aktuellen Wissensstand von Lernenden zu erfassen, bedarf es einer Re-Repräsentation des Wissens. Ein Schwerpunkt dieser Arbeit liegt dabei auf der Untersuchung re-repräsentierten Wissens.

2.2.2 Lehrerexpertise

Das folgende Kapitel fokussiert die Expertise von Lehrkräften. Die diagnostische Expertise legt dabei das Fundament für eine objektive, reliable sowie valide Bewertung von Lernergebnissen vor. Reinisch (2009) beleuchtet das Konstrukt der

Lehrerprofessionalisierung aus soziologischer wie auch wissenspsychologischer Perspektive. Aus soziologischer Sicht umfasst „Professionalisierung“ die Ausbildungszeit, die als Ziel die erfolgreiche Befähigung eines beabsichtigten Berufs hat (ebd., S. 34; Mieg, 2005). Lehrkräfte werden von Etzioni (1969) allerdings als „Semi-Professionals“ verstanden (Reinisch, 2009, S. 36). Dieser Zuordnung zufolge lassen sich die Merkmale eines Lehrberufs keinen professionellen Berufen zuordnen. Dies impliziert, dass Lehrkräfte nie den Stand von Professionellen erhalten können. Oevermann (1996) verweist dabei auf die Professionalisierungsbedürftigkeit von Lehrkräften (vgl. Reinisch, 2009, S. 35). Die soziologische Perspektive betrachtet Lehrerprofessionalität zwar als notwendig, doch per Definition jedoch nur „semi“ erreicht werden. Lehrkräfte können dieser Argumentationsweise nach nie als Experten in ihrem Gegenstandsbereich angesehen werden.

Aus wissenspsychologischer Perspektive zeichnet sich ein Experte durch Professionalität aus (Reinisch, 2009, S. 37). Dies basiert auf der kognitionspsychologischen Grundlage der Expertiseforschung. Der englischsprachige Raum verwendet den Begriff der „expert teachers“ wobei im deutschsprachigen Raum zwischen unterschiedlichen Expertisegraden unterschieden wird (vgl. Gruber, 1994, Ifenthaler, 2006, Pirnay-Dummer, 2006). Kennzeichnend für den deutschsprachigen Raum ist die Verwendung des Lehrers als Experten nach Bromme (1992). Bromme (2008) unterscheidet bei Lehrkräften in Anlehnung an Shulman (1986) zwischen folgenden Wissenstypen: dem Inhaltswissen, dem curricularem Wissen, der Philosophie des Schulfachs, dem pädagogischen Wissen und dem fachspezifisch-pädagogischen Wissen (vgl. Bromme & Haag, 2004, S. 809; Bromme, Rheinberg, Minsel, Winteler & Weidenmann, 2006, S. 314). Ersteres umfasst das domänenspezifische Wissen aus dem jeweiligen Unterrichtsfach. Das curriculare Wissen beinhaltet zudem die Bildungsziele. Das fachspezifisch-pädagogische Wissen umfasst didaktisches Wissen in Bezug auf die Durchführung von Unterricht. Für eine präzise Diagnose von Lernergebnissen setzt die vorliegende Arbeit insbesondere die ersten beiden Wissenstypen voraus. Shulman (2000) verweist auf Studien, die im Bereich der domänenabhängigen Expertise sowie dem pädagogischen Wissen durchgeführt wurden.

Besser & Krauss (2009) verweisen auf die Notwendigkeit, „den Ansatz der Expertiseforschung aus der kognitiven Psychologen in den Bereich der Lehrerforschung zu übertragen“ (ebd., S. 80). Demzufolge können Lehrkräfte aus wissenspsychologischer Perspektive als Experten in ihrer Domäne angesehen werden. Die vorliegende Forschungsarbeit stützt sich auf diesen Ansatz und fokussiert die diagnostische Expertise, welche im nächsten Abschnitt dargestellt wird.

2.2.3 Diagnostische Expertise

Die diagnostische Expertise spielt für die Leistungsfeststellung eine entscheidende Rolle. Um Prüfungsleistungen hinreichend objektiv, reliabel und valide zu erfassen und zu bewerten müssen Hochschuldozenten und Lehrkräfte diese Expertise vorweisen. Der folgende Abschnitt präzisiert dieses Konstrukt und bettet es in den empirischen Forschungsstand ein.

Helmke (2009) definiert das Konstrukt der diagnostischen Expertise als das Wissen über eine objektive, reliable und valide Bewertung sowie die Kenntnis über mögliche Fehlerquellen, die in den Prozess der Bewertung einfließen. Schließlich umfasst es die Fähigkeit einer Lehrkraft, nach Tests zu recherchieren, sowie selbst einen Test zu entwickeln, anzuwenden und auszuwerten (vgl. Helmke, 2009, S.122 f). Schrader & Helmke (2002) beschreiben dieses Konstrukt als eine „Fähigkeit Schülermerkmale und Aufgabenschwierigkeit zutreffend einzuschätzen“ (ebd., S. 48). Sie umfassen sowohl die diagnostische Expertise als auch die diagnostische Fähigkeit. Die erste fokussiert die Fähigkeit, unterschiedliche Aufgabenniveaus zu berücksichtigen. Die zweite fokussiert das Beobachten sowie den Einsatz geeigneter Diagnoseinstrumente (ebd, S. 48). Spinath (2005) verweist auf die Schwierigkeit, das Konstrukt der diagnostischen Expertise überhaupt messbar zu machen (ebd., S. 93). Die vorliegende Arbeit orientiert sich an den theoretischen Annahmen der diagnostischen Expertise (nach Helmke, 2009) und verzichtet auf die Verwendung des Begriffs der diagnostischen Kompetenz.

Helmke (2009) unterscheidet drei Komponenten der Urteilsgenauigkeit: die Niveaukomponente, die Streuungskomponente, sowie die Rang-ordnungskomponente. Erstere gibt an, wie gut die Lehrkraft das Niveau ihrer

Klasse einschätzen kann. Sie lässt Rückschlüsse darauf ziehen, inwiefern der Lehrer seine Schüler über- bzw. unterbewertet. Die Streuungskomponente gibt an, wie genau der Lehrer die Streuung der Noten innerhalb der Klasse einschätzen kann. Die Rangkomponente zeigt, wie gut die Lehrkraft die Schülerleistungen in Ränge ordnen kann.

Jäger (2009) gibt einen Überblick über die diagnostischen Aufgaben des Lehrers im Kontext des Unterrichtens und zwar über verschiedene Wissensbereiche, welche die diagnostische Expertise beim Unterrichten umfasst (ebd., 2009). Dabei differenziert er zwischen folgenden Wissensbereichen: 1) das Kompetenzwissen, 2) das Bedingungswissen, 3) das technologische Wissen, 4) das Änderungswissen, und 5) das Vergleichswissen. Das Kompetenzwissen beantwortet die Frage, wie etwas diagnostiziert werden kann. Ob die Lehrkraft sich bewusst ist, welche Möglichkeiten (Experten, Instrumente) ihr zur Verfügung stehen um bestimmte Schülermerkmale zu identifizieren. Das Bedingungswissen fokussiert die „möglichen Bedingungen eines gegebenen Verhaltens“ (ebd., S. 475). Das technologische Wissen umfasst das Wissen über diagnostische Grundlagen, in welcher Weise sich etwas so objektiv wie nur möglich erfassen lässt. Das Änderungswissen fokussiert das Wissen darüber, wie sich Lernprozesse optimieren lassen. Das Vergleichswissen fokussiert das Einordnen einer bestimmten Schülerleistung in den Gesamtkontext - beispielsweise einer ganzen Lernergruppe.

2.2.3.1 Empirischer Forschungsstand

Der empirische Forschungsstand zeigt, dass es Lehrkräften recht gut gelingt, die Schülerleistungen innerhalb einer Klasse in Ränge zu ordnen (vgl. Schrader & Helmke, 2002, S. 50). Außerdem fällt es in heterogenen Lernergruppen leichter, Schülerleistungen zu bewerten als in homogenen (ebd., S. 48). Dabei beeinflusst die Klassengröße Wild & Rost (1995) zufolge nicht den Präzisionsgrad des Lehrerurteils in Hinsicht auf die Schülerleistungen.

McElvany et al. (2009) untersuchten, wie präzise Lehrkräfte den Schwierigkeitsgrad von Unterrichtsmaterialien mit Texten und instruktionalen Bildern einschätzten. Dabei zeigte sich, dass diese durch die Lehrkräfte weniger akkurat eingeschätzt wurden als durch die vorhergehenden Befunde von denen Hoge & Colodarci (1989) berichtet. Aufgaben, die vom Schwierigkeitsgrad eher

leicht und homogen waren, konnten durch die Lehrkräfte besser unterschieden werden.

Dies verdeutlicht, dass Leistungen in kontrastreichem Ergebniskontext einfacher eingeschätzt werden können als im ähnlich leistungsstarken Kontext. Aufgabenschwierigkeiten hingegen können einfacher hinsichtlich des Schwierigkeitsgrades eingeschätzt werden, wenn sie ähnlich leicht im Vergleich sind. Die Berufsdauer hatte dabei keinen Einfluss auf die Genauigkeit der Einschätzungen (vgl. McElvany et al., 2009). Diese findet auch in der vorliegenden Forschungsarbeit keine Berücksichtigung.

Krolak-Schwerd, Böhmer & Gräsel (2012) zeigten, dass Laien im Gegensatz zu Lehrkräften Schülerleistungen unsystematisch beurteilten. Diese Studie verglich die Leistungsbeurteilungen von N = 50 Gymnasiallehrern mit N = 48 Studierenden, welche keine pädagogische Erfahrungen hatten. Dabei erhielten diese unterschiedliche Vorinformationen über die einzelnen Schüler, sowie über deren soziales Umfeld. Die Laien berücksichtigen weder die individuelle, noch die soziale Bezugsnorm als Vergleichsmaß (vgl. ebd. S. 120). Dies verdeutlicht, dass Lehrkräfte auf Grundlage diagnostischer Expertise auf unterschiedliche Bezugsnormen zurückgreifen und diese in den Prozess der Bewertung einfließen lassen können.

Spinath (2005) argumentiert gegen die Annahme des Konstrukts der diagnostischen Expertise. Ihre empirische Studie umfasste N = 723 Schüler und N = 43 Lehrer an insgesamt vier Grundschulpopulationen. Sie untersuchte die drei Komponenten der Urteilsgenauigkeit und fand dabei nur eine geringe Genauigkeit der Lehrerurteile bezüglich deren Schülermerkmale. Die empirischen Befunde der hier dargestellten Untersuchungen legen eine unzureichende Diagnosefähigkeit von Lehrkräften bezüglich der Bewertung von Schülerleistungen nahe.

Boud & Falchikov (1995) geben einen Überblick an empirischen Untersuchungen, die die Selbsteinschätzung der Lernenden mit der Einschätzung der Lehrkräfte vergleicht. Dabei unterschätzen sich Lerner, die als sehr gut eingeschätzt wurden häufig, wohingegen sich leistungsschwache Schüler eher überschätzten (ebd., 1995). Die vorliegende Arbeit berücksichtigt nicht die Selbsteinschätzung der Lernenden, sondern fokussiert ausschließlich die Bewertungen der Lehrkräfte, die laut Definition (2.4.1) als Experten in ihrem Gegenstandsbereich eingestuft werden. Pirnay-Dummer (2012) weist darauf hin, dass Lernende versuchen, sich

bei der schriftbasierten Wissensexternalisierung am Erwartungshorizont des Leistungsbeurteilers zu orientieren (ebd., S. 80). Die Ergebnisse der Hamburger Aufsatzstudie (Hartmann & Lehmann, 1989) verdeutlichen, dass Schüler ihre Aufsätze inhaltlich so aufbauten, dass sie dem Meinungsbild des Lehrers entsprechen (vgl. Schrader & Helmke, 2002, S. 241).

Auch dieser Effekt wird in der vorliegenden Forschungsarbeit nicht berücksichtigt.

Die Ergebnisse der erwähnten Studien deuten darauf hin, dass sich der diagnostische Expertisegrad von Lehrkräften auf die Präzision der Leistungsbeurteilung auswirkt. Dies lässt vermuten, dass Lehrkräfte deren diagnostische Expertise gezielt geschult wird, eine objektivere, reliablere sowie validere Feststellung von Lernergebnissen vornehmen. Die offensichtlich optimierbare Expertise von Lehrkräften hinsichtlich der Diagnose deutet auf die Notwendigkeit hin, Lehrer bei der Bewertung textbasierter Schülerleistungen durch ein geeignetes (wissenschaftlich fundiertes) Hilfsmittel zu unterstützen.

3 Pädagogische Diagnostik

Die pädagogische Diagnostik befasst sich mit der Analyse von Lernvoraussetzungen, Lernprozessen sowie Lernergebnissen und Lernumgebungen (vgl. Schwarzer, 1982, S. 6; Ingenkamp & Lissmann, 2008, S. 13). Die vorliegende Arbeit fokussiert die Bewertung von Lernergebnissen. Dies erfordert ein Wissen bei Lehrkräften, das externalisierte Wissen von Lernenden hinreichend objektiv, zuverlässig sowie gültig zu erfassen sowie einzuschätzen (vgl. Schrader, 2006). Der aktuelle Stand der Forschung zeigt, dass dies insbesondere in den schriftlichen Prüfungssituationen nicht gegeben ist (vgl. Klauer, 1978; Ingenkamp & Lissmann, 2008, Linn, Klein & Hart, 1970).
Die folgenden Kapitel fokussieren die Leistungsdiagnostik durch Lehrende und daran anschließend die technologiegestützte Leistungsdiagnostik. Dabei wird jeweils der aktuelle Forschungsstand dargestellt.

3.1 Leistungsdiagnostik durch Lehrende

Leistungsmessung und - beurteilung spielen im Schul- und Hochschulkontext eine wesentliche Rolle (vgl. Schwarzer, 1982, S. 10). Legt man die Definition der pädagogischen Diagnostik zugrunde (die im vorherigen Kapitel beschrieben wurde), so lässt sich daraus schließen, dass die Aufgabe von Lehrkräften mitunter die präzise Leistungsfeststellung umfasst. Um eine präzise Leistungsfeststellung zu vollziehen, müssen bestimmte Standards (gemäß der Objektivität, Reliabilität sowie Validität) erfüllt sein. Diese finden sich jedoch häufig nicht vor, wenn es zu der Analyse von textbasierten Prüfungsleistungen kommt. Das vorliegende Kapitel unterteilt sich in zwei größere Bereiche der Leistungsdiagnostik. Zunächst folgen eine Eingrenzung der Begrifflichkeit und anschließend eine Darstellung des aktuellen Forschungsstandes auf Grundlage empirischer Untersuchungen.

3.1.1 Präzision der lehrerbasierten Leistungsdiagnostik

3.1.1.1 Leistungsmessung und Leistungsbeurteilung

Ingenkamp (1981) differenziert zwischen Leistungsmessung und Leistungsbeurteilung. Bevor es zu einer präzisen Beurteilung der externalisierten Wissensstrukturen kommen kann, bedarf es einer geeigneten Problemstellung, die

die Lernenden initiiert, ihr Wissen in einer ihnen gewohnten und geeigneten Weise abzubilden. Leistungsmessung umfasst damit die initiierten Externalisierungsprozesse von Wissen, welche der Leistungsbeurteilung vorausgeht. Die Leistungsbewertung beinhaltet die Überprüfung sowie Rückmeldung an den Lernenden, wie weit sein Wissen aktuell noch vom Lernziel entfernt liegt. Um diesen Soll-Ist-Vergleich vornehmen zu können, bedarf es eines aus den Lernzielen abgeleiteten Erwartungshorizontes, welcher durch die Bewertungskriterien operationalisiert und dadurch überprüfbar gemacht wird. Diese daraus abgeleiteten Kriterien entscheiden bei der Leistungsbewertung, wann eine Leistung als gut eingeschätzt wird. Dabei können verschiedene Bezugsnormen als Vergleich der einzelnen Schülerleistungen herangezogen werden (vgl. Rheinberg, 2002, S. 59; Rheinberg & Fries, 2010).

Schrader und Helmke (2002) unterscheiden zwischen expliziten und impliziten Einschätzungen von Lehrkräften in der Leistungsbewertung (ebd., S. 45 f). Explizite Einschätzungen beziehen sich beispielsweise auf schriftliche oder mündliche Leistungskontrollen, die dem Schüler auf Grundlage einer der Bezugsnormen gezielt zu einem bestimmten Themengebiet Rückmeldung geben. Danach folgen dieser die anschließende Leistungsbewertung (ebd., S. 45). Implizite Einschätzungen umfassen eher eine informelle Rückmeldung, die die Lehrkraft dem Lernenden gibt und erfordern einen wesentlich verkürzten Prozess der Urteilsbildung, da sie unmittelbar im Geschehen - beispielsweise im Unterricht - erfolgen (ebd., S. 46). Rheinberg (1978) unterscheidet zwischen *informeller* sowie *formeller* Leistungsdiagnostik. Die Begrifflichkeiten können synonym zu den impliziten- als auch expliziten Einschätzungen herangezogen werden. Sowohl implizite als auch informelle Bewertungen berücksichtigen weniger die Gütekriterien, wohingegen explizite und formelle Bewertungen dies zu integrieren versuchen. Die vorliegende Forschungsarbeit fokussiert die expliziten-, formellen Bewertungen von Lernergebnissen am Beispiel von schriftlichen Prüfungen. Dabei liegt ein besonderes Augenmerk auf der von Lehrkräften Kriterien orientierten Bewertung textbasierter Lernergebnisse. Das Schreiben von Schülertexten hat primär die Überprüfung bereits erreichter Lernziele zum Ziel sowie die daran anschließende Bewertung (vgl. Eigler, 1997, S. 366).

3.1.1.2 Funktion von Schule

Ingenkamp & Lissmann (2008) definieren Schulleistung als „die von der Schule initiierten Lernprozesse und Lernergebnisse der Schüler“ (ebd., S. 131). Die Schule hat nach Fend (2008) die Aufgabe Lernende zu qualifizieren, zu selektieren (allokieren) sowie zu legitimieren (ebd., S. 50). Die Schule steht somit für einen Ort des Lernens, welcher immer beabsichtigt, Schüler auf einen höheren Expertisegrad zu befördern. Die Selektionsfunktion fokussiert das Feststellen des aktuellen Lernstandes von Schülern, um diese geeigneten Lerngruppen zuzuordnen. Damit diese Zuweisung zu einer bestimmten Lerngruppe erfolgen kann, bedarf es einer präzisen Diagnose von Lernprozessen sowie Lernergebnissen, um diese hinreichend zu fördern und zu optimieren. Die Legitimationsfunktion findet durch das Erfassen sowie Bewerten von Schulleistungen statt, die den einzelnen Lernenden hinsichtlich verschiedener Ausbildungsplätze legitimieren. Hierfür sind geeignete Messinstrumente bezüglich der Erfassung sowie der Auswertung notwendig, um den Lernprozess sowie den Leistungsstand zu veranschaulichen. Im Bildungskontext wird zwischen verschiedenen Ebenen unterschieden: der Mikro-, der Meso-, sowie der Makroebene (vgl. Cortina, 2006). Die Mikroebene fokussiert den Lerner, die Mesoebene umfasst die Klassen sowie die Schulebene und die Makroebene zielt auf das Schulsystem. Auf der Makroebene, welche die selektive Funktion einschließt, finden die Diskussionen in Bezug auf die Chancengleichheit statt (ebd., S. 495).

3.1.1.3 Bezugsnormorientierung

Die Bezugsnormen unterteilen sich in der sachlichen-, der individuellen- sowie der sozialen Bezugsnorm (vgl. Rheinberg, 2002, 2006, 2008, 2009; Rheinberg & Fries, 2010). Dabei entscheidet der Zweck der Rückmeldung, welche Bezugsnorm herangezogen wird. Neben der sozialen (klasseninternen) Bezugsnorm unterscheidet Rheinberg (2006) zwischen der kriterialen- und der individuellen Bezugsnorm. Während die Kriterien geleitete Bewertung aufzeigt, wieweit der Lerner vom Lernziel entfernt ist und zudem eine Vergleichbarkeit ermöglicht, zeigt die individuelle Bezugsnorm vor allem individuelle Lernprozesse auf. Rheinberg (2002) verweist auf die „blinden Flecken“ im Zusammenhang mit den

Bezugsnormen (ebd., S. 64, 2009). Nachdem die Schülerleistungen mit einer Bezugsnorm bewertet wurden, kommt es zur eigentlichen Notengebung.

3.1.1.4 Funktion der Notengebung

Die Notengebung erfolgt, nachdem die Leistungen auf Grundlage der zuvor festgelegten Kriterien bewertet wurden. Diese verfolgen unterschiedliche Zielsetzungen. So kann es passieren, dass der Prozess der Leistungserfassung sowie -bewertung zwar objektiv, zuverlässig und valide erfolgte, die eigentliche Notengebung allerdings den wissenschaftlichen Standards nicht mehr genügte. Weinert & Schrader (1986) begründen diese Problematik in den unterschiedlichen Zielsetzungen die durch die Notengebung angestrebt werden (ebd., S. 16 f). Leistungsrückmeldungen erfolgen zumeist in Form von Noten, welche pädagogische - sowie gesellschaftliche Funktionen haben (vgl. Tent, 2006; Tent & Birkel, 2010). Die *pädagogische Funktion* umfasst in erster Linie die Sozialisation, die Rückmeldung sowie die Motivation. Dabei soll der Schüler für sein weiteres Lernverhalten motiviert werden, indem er Rückmeldung über seinen aktuellen Lernprozess erhält. Die *gesellschaftliche Funktion* umfasst die Berechtigung, die Selektion sowie die Kontrolle. Diese weisen den Lerner einer bestimmten Gruppe zu und ermöglichen ihm spätere Ausbildungs- sowie Arbeitsplätze. Schließlich fokussieren sie eine Transparenz schulpolitischer sowie pädagogischer Maßnahmen (vgl. Tent, 2006, S. 873).

3.1.1.5 Einfluss von Fehlerquellen auf die Leistungsbewertung

Der Bewertungsprozess von Schülerleistungen wird von verschiedenen Fehlern beeinflusst. Dabei werden verschiedene Fehlerquellen unterschieden: die Tendenz zur Mitte,- zu extremen Urteilen,- zur Milde, der Pygmalioneffekt, der Referenzfehler oder beispielsweise der Halo-Effekt. Diese Störungen beeinflussen die Genauigkeit, mit der eine bestimmte Leistung eingeschätzt wird (für einen Überblick siehe Helmke, 2009, S. 138; Bohl, 2005, S. 66; Ziegenspeck, 1999). Schrader und Helmke (2002) deuten auf die Problematik hin, dass die Voreingenommenheit und Erwartungen des Beurteilers den Bewertungsprozess verzerrt (ebd., S. 46 f). Weinert (2002) weist auf die Notwendigkeit hin, wissenschaftlich fundierte Instrumente in der schulischen Leistungsmessung

heranzuziehen (ebd., S. 359). Hierfür soll die vorliegende Arbeit einen Beitrag leisten.

Das Heranziehen der klasseninternen Bezugsnorm als Indikator für die Einschätzung des Leistungsstandes eliminiert einen Vergleich mit weiteren Lerngruppen, die nicht Bestandteil dieses Klassengefüges sind (vgl. Rheinberg, 2002; Ingenkamp, 1989). Die Konsequenz hieraus ist, dass eine bestimmte Schülerleistung entweder als „gut“ oder „ausreichend“ bewertet wird, je nachdem ob sich der Lerner in einer leistungsstarken oder leistungsschwachen Klasse befindet. Schließlich können beim Heranziehen der sozialen Bezugsnorm individuelle Lernzuwächse oder auch die Lernzuwächse innerhalb einer bestimmten Lerngruppe nicht sichtbar gemacht werden. Das Heranziehen der individuellen Bezugsnorm verdeckt hingegen die individuellen Entwicklungsverläufe zwischen verschiedenen Fachbereichen und verdeckt den Vergleich zur Lernergruppe, als auch dem kriterial orientierten Erwartungshorizont. Das Heranziehen der sachlichen Bezugsnorm verdeckt hingegen individuelle Lernfortschritte.

3.1.2 Empirischer Forschungsstand

Der aktuelle Forschungsstand verdeutlicht die Problematik der Bewertung textbasierter Leistungen. Ingenkamp & Lissmann (2008) geben einen Überblick über die empirischen Untersuchungen zu den Gütekriterien schriftlicher Arbeiten. Dabei werden im Folgenden einzelne Studien in Bezug auf die Objektivität, die Reliabilität sowie die Validität exemplarisch dargestellt.

3.1.2.1 Problematik der Gütekriterien

Der aktuelle Stand zeigt, dass sich Gutachter bei der Analyse von textbasierten Schülerleistungen in ihren Urteilen unterscheiden. Dies verletzt das Kriterium der Objektivität. Hierzu rezipierte Birkel & Birkel (2002) eine Studie von Weiss (1995), indem sie N = 89 Grundschullehrer beim Bewerten von Schüleraufsätzen untersuchten. Sie analysierten je zwei Versionen von insgesamt vier Aufsätzen, welche sich in der Qualität, der Länge, sowie der Anzahl an Rechtschreibfehlern unterschieden. Die Einschätzungen der Lehrkräfte unterschieden sich zwischen

drei bis vier Notenstufen. Dies verdeutlicht eine Verletzung der Auswertungsobjektivität.

Auch die empirischen Befunde von Ingenkamp (1995) deuten auf eine mangelnde Auswertungsobjekivität hin. Er untersuchte (N = 37) Grundschulklassen des sechsten Schuljahres in Berlin mithilfe eines validen Mathematiktests. Die erzielten Gesamtpunkte und die von den Lehrkräften vergebenen Noten wichen zwischen den Klassen stark voneinander ab, was wiederum die Problematik der sozialen Bezugsnormorientierung verdeutlicht. Ob sich ein Schüler in einer leistungsstarken- oder leistungsschwachen Klasse befindet, beeinflusst die Bewertung seiner Prüfungsleistung. Dabei orientieren sich Lehrer gerne an der klasseninternen Bezugsnorm (vgl. Ingenkamp & Lissmann, 2008, S. 146; Schrader & Helmke, 2002, S. 50). Das heißt, sie legen den Leistungsstand der Klasse zugrunde, um die Noten zu bestimmen. Demzufolge erfolgt eine mangelhafte Auswertungsübereinstimmung in zwei wesentlichen Bereichen der Leistungsfeststellung. Erstens bei der Punkteermittlung, und zweitens bei der Notenfeststellung. Die Gewichtungen, die nach dem Feststellen der Punktezahl erfolgt, entscheiden letztendlich über die vergebene Note. Ebenso entscheidet das Heranziehen der Bezugsnorm, welche Note letztendlich bestimmt wird.

Die folgende Untersuchung, welche exemplarisch dargestellt wird, zeigt die Unzuverlässigkeit von Lehrerurteilen. Eels (1930; 1995) untersuchte die Leistungsfeststellung von 61 Lehrkräften, welchen er dieselben Kurzaufsätze in den Fächern Geografie sowie Geschichte nach elf Wochen erneut bewerten ließ. Dabei ergaben die statistischen Analysen einen Korrelationskoeffizient von r = 0.25. Dies lässt darauf schließen, dass bereits die zugrunde gelegten Kriterien nicht messgenau waren. Zudem fand sich keine Bewertungskonsistenz zwischen den Lehrkräften, d. h. die Notengebung erfolgte nicht unabhängig vom Bewerter, sodass es für den Schüler entscheidend war, welche Lehrkraft die eigenen Lernergebnisse analysierte.

In Bezug auf die Gültigkeit von Lehrereinschätzungen zeigten die Ergebnisse einer Metaanalyse bei der *prognostischen Validität* einzelner sowie gemittelter Schulnoten auf die späteren Studiennoten, mittlere Validitätskennwerte von $\rho = 0.26$ bis $\rho = 0.53$ (vgl. Trapmann, Hell, Weigand & Schuler; 2007; Schuler, 2010, S. 601). Insgesamt betrachtet, stellen Noten einen unzureichenden Prädiktor dar. Schuler (2010) gibt einen Überblick an empirischen Untersuchungen im

Bereich der prognostischen Validität von Noten. Die Untersuchungen deuten auf eine invalide Güte von Notengebungen hin. Unklar bleibt, was genau die Noten zeigen, da sie ein Sammelspektrum an bisherigen Leistungsfeststellungen abbilden und weder der Prozess noch die Kriterien transparent sind, die hierzu geführt haben, noch welche Gewichtungen bei den Leistungen herangezogen wurden und welche Bezugsnorm im Einzelnen verwendet wurde. Dies verhindert einen Leistungsvergleich, der den Studienerfolg vorhersagen soll. Auch der interindividuelle Unterschied ist somit nicht möglich.

3.1.2.2 Fehlerquellen

Bei der Bewertung textbasierter Schülerleistungen beeinflussen folgende Faktoren die Notengebung: die Textlänge, die Rechtschreibfehler, die Handschrift, die Reihenfolge in der die Leistungen bewertet werden, sowie die Beliebtheit der Schüler (für einen Überblick siehe Ingenkamp & Lissmann, 2008).

Pohlmann & Möller (2007) analysierten den Primingeffekt auf die Beurteilung. Sie untersuchten die Assimilations- und Kontrasteffekte bei der Bewertung von Texten. Dabei ließen sie N = 40 Lehramtsstudierende die Qualität von jeweils zwei Buchrezensionen einschätzen. Die eine Hälfte der Probanden schulten sie daraufhin, Gemeinsamkeiten in Texten zu suchen die andere Hälfte an Versuchspersonen schulten sie, Unterschiede im Text zu suchen. Die Ergebnisse dieser Untersuchungen zeigen, dass Texte als ähnlicher eingeschätzt wurden, wenn vorher ein Priming hinsichtlich der Ähnlichkeit erfolgte. Dies verdeutlicht die Problematik des Reihenfolgeneffektes bei der Leistungsfeststellung.

Auch im Bereich alternativer Bewertungsinstrumente (wie beispielsweise das Portfolio) finden sich Studien, die insbesondere die Leistungsbewertung fokussieren. Gläser-Zikuda, Rohde & Schlomske (2010) geben hierfür einen Überblick. Typische Vertreter, die sich mit alternativen Bewertungsinstrumenten beschäftigten sind beispielsweise Winter (2008) und Bohl (2005).

Die vorliegende Arbeit fokussiert ausschließlich klassische Bewertungsinstrumente (in Form von textbasierten Klausuren), da die Ergebnisse der empirischen Untersuchungen, die oben bereits angedeutet wurden, zeigen, wie schwierig es ist, insbesondere textbasierte Schülerleistungen mit genügend Übereinstimmung zu bewerten. Die vorliegende Studie fokussiert dabei vor allem die Bewertung von

textbasierten Leistungen. Dies ist zum einen im Instrument begründet, welches in Kapitel 4 beschrieben wird. Zum anderen dominieren textbasierte Leistungen, beispielsweise in Form von Klausuren im Schul- und Hochschulkontext. Außerdem sind textbasierte Schülerleistungen von den Gütekriterien her betrachtet am unzureichendsten bewertet, obwohl sie in den meisten Unterrichtsfächern ein beliebtes Externalisierungsformat darstellen.

3.2 Technologiegestützte Leistungsdiagnostik

Pirnay-Dummer (2012) charakterisiert Technologien im pädagogischen Kontext als eine Unterstützung für Lehrkräfte, welche auf Lehr- Lerntheoretischen Ansätzen beruht (ebd., S. 6). Eckert (1998) verweist ebenso auf die Dringlichkeit, theoriegeleitete Verfahren bei der Erfassung von Wissen zu verwenden (ebd., S. 16). Wie bereits in der theoretischen Herleitung (1.2.3) verdeutlicht, kann intern abgespeichertes Wissen nie direkt diagnostiziert werden, sondern immer nur eine realisierte Abbildung davon. Aus kognitionswissenschaftlicher Sicht sind bei der Wissensdiagnose insbesondere die abgebildeten Modellstrukturen interessant (ebd., S. 11). Damit wird vorausgesetzt, dass das kognitive System sein Wissen abrufen und externalisieren kann. Dabei entscheiden die kognitiven Anforderungen, die das Messinstrument an den Externalisierungsprozess von Modellstrukturen stellt, inwieweit diese letztendlich abgebildet werden können (ebd., S. 18).

3.2.1 Empirischer Forschungsstand

Der folgende Teil der Arbeit fokussiert die technologiegestützte Analyse textbasierter Wissensrepräsentationen. Dabei folgt die Darstellung des aktuellen Forschungsstandes innovativer Technologien im Bereich der Leistungsfeststellung. Wie bereits in Kapitel 3.1 gezeigt wurde, weisen die empirischen Untersuchungen zu den Gütekriterien bei der Bewertung von schriftlichen Arbeiten nur geringe Kennwerte auf. Lenhard, Baier, Hoffmann & Schneider (2007) gibt einen Überblick an Untersuchungen, die überwiegend im englischsprachigen Raum stattgefunden haben, um die Bewertung von Aufsätzen zu automatisieren (ebd., S. 156). Hierfür bedarf es immer eines durch Experten festgelegten Außenkriteriums als Referenzmodell. Die Bestimmung eines eindeutig festgelegten

schriftlichen Außenkriteriums ist dabei immer schwierig (vgl. Bühner, 2004, S. 61, Lenhard et al., 2007, S. 155).

Koul, Clariana & Salehi (2005) verglichen die menschliche Bewertung von N = 22 Essays mit den mittels LSA (*Latent Semantic Analysis*) und mittels *ALA-Reader* (Analysis of Lexical Aggregates) ermittelten Bewertungen. Dabei korrelierten die menschliche Einschätzung der Essays mit den mittels ALA-Reader ermittelten Werten mit $\rho = 0.60$. Der *ALA-Reader* wurde von Clariana (2004) entwickelt und verlangt die Texte mittels durch Experten festgelegter Begrifflichkeiten. Clariana (2010) kommt zu der Schlussfolgerung, dass sich manche Gegenstandsbereiche besser eignen, um mithilfe des ALA-Readers analysiert zu werden. Je spezifischer die in den Essays verwendeten Begrifflichkeiten, desto besser können die Bewertungen mithilfe des ALA-Readers bestimmt werden (ebd. S. 127).

Lenhard et al. (2007) untersuchten die automatisierte Leistungsfeststellung am Beispiel von LSA. Im Fokus ihrer Studie stand die Überprüfung, inwiefern sich die Bewertung offener Antworten mittels LSA ermitteln lässt. Dabei fanden sich hohe Übereinstimmungen der LSA- Auswertungen mit denen durch Experten vergebenen Punkte.

Die empirischen Befunde von Schlomske & Pirnay-Dummer (2009) deuten darauf hin, dass sich lernerabhängige Veränderungen nahezu funktional vorhersagen lassen. Dabei zogen sie als Außenkriterien sprachlich orientierte, externalisierte Modelle von Experten sowie von fortgeschrittenen Lernern heran, um die Entwicklungen in einer Lernergruppe vom Anfänger zum fortgeschrittenen Lerner aufzuzeigen. Die Ergebnisse deuten darauf hin, dass mithilfe von Außenkriterien die Entwicklung einzelner Lerner technologiebasiert vorhergesagt werden kann. Dies führt zu der Überlegung, dass sich Leistungsfeststellungen mithilfe technologiebasierter Instrumente unterstützten lassen.

Pirnay-Dummer & Ifenthaler (2010) weisen darauf hin, dass Lernende von technologieunterstützten Diagnoseinstrumenten profitieren, da ihnen diese eine unmittelbare Rückmeldung im Sinne der Selbstbeurteilung ermöglichen (ebd, S. 77). Inwiefern sich eine zeitnahe Rückmeldung des in Textform externalisierten Wissens günstig auf die Schülerleistungen auswirkt, verdeutlicht die Hamburger Aufsatzstudie von Hartmann & Lehmann (1989). Wohingegen sich eine zeitgleiche Rückmeldung inmitten des Externalisierungsprozesses eher hinderlich auf die Leistung auswirkt (vgl. Helmke & Schrader, 2002, S. 242). Da die klassische Form

der Aufsatzbewertung sowie textbasierter Schülerleistungen üblicherweise immense Zeit von Seiten der Lehrkräfte erfordert, wäre ein auf die Gütekriterien hin geprüftes, automatisiertes Instrument eine Unterstützung um Lernende schnellere Rückmeldung auf ihren aktuellen Leistungsstand zu geben.
Pirnay-Dummer (2012) untersuchte, inwiefern sich Textvergleichsindizes eignen, um die Leistungsermittlung von Hausarbeiten im Hochschulbereich vorherzusagen. Dabei untersuchte er in einer Pilotstudie die in Form von Noten bewerteten Hausaufgaben in den Fächern Erziehungswissenschaft und Wirtschaftswissenschaft. Als Vergleichsmaßstab zog er die von den einzelnen Studierenden primär als Grundlage verwendete Literatur heran. Die Ergebnisse zeigten, dass sich die Noten auf Grundlage der Vergleichsindexe im Fachbereich Volkswirtschaftslehre sowohl auf struktureller- als auch auf semantischer Ebene eigneten, um die Note mit einer hohen Wahrscheinlichkeit vorherzusagen. Im Fachbereich Erziehungswissenschaft fanden sich solche Vorhersagewerte jedoch nicht (vgl. Pirnay-Dummer, 2012).

3.2.2 Methodologische Anmerkungen

Der Vergleich einzelner textbasierter Prüfungsleistungen mittels LSA verlangt große Datensätze (vgl. Landauer, Foltz & Laham, 1998; Pirnay-Dummer & Walter, 2009), was im schulischen Kontext (kleine Klassengrößen sowie kürzere Texte) als Prüfungsleistung oft nicht gegeben ist.
Der aktuelle Forschungsstand hat gezeigt, dass es wenige technologiegestütze Verfahren gibt, um die Analyse textbasierter Lernergebnisse zu unterstützen. Es wurde gezeigt, dass insbesondere im schulischen Kontext, wo sich kleine Klassengrößen finden und wo es zu schnellen Leistungsurteilen kommen muss, geeignete Verfahren fehlen. Dem geht die vorliegende Forschungsarbeit nach und untersucht ein zeitökonomisches Instrument welches hinsichtlich der wissenschaftlichen Gütekriterien gute Kennwerte aufweist. Das folgende Kapitel stellt dieses Instrument, welches aus den Gegenüberstellungen der vorhandenen Instrumente als geeignet festgestellt wurde, näher dar.

4 Instrumente und methodologische Diskussion

Das folgende Kapitel stellt die Instrumente T-MITOCAR (Text-Model Inspection Trace oft Concepts and Relations) und AKOVIA (Automated Knowledge Visualisation and Assessment) sowie die Qualitative Inhaltsanalyse vor, die in der vorliegenden Arbeit herangezogen wurden. Die ersten beiden Technologien fokussieren die Analyse externalisierten Wissens in Form von Texten. Letzteres analysiert Texte systematisch, indem es auf Grundlage von zuvor festgelegten Regelwerken Kategorien bildet. Am Schluss erfolgt eine methodologisch-kritische Diskussion.

4.1 T-MITOCAR

T-MITOCAR ist eine computerbasierte Software, die von Pirnay-Dummer entwickelt wurde und basiert auf der Theorie der mentalen Modellbildung (ebd., 2007, 2010, 2011, 2012). Es ist eine Weiterentwicklung von MITOCAR (Model Inspection Trace of Concepts and Relations), welches von Pirnay-Dummer (2006, 2010) zur Wissensdiagnose und Abgrenzung unterschiedlicher Expertisegrade entwickelt wurde. MITOCAR analysiert die Expertisegrade kleinerer Gruppen aus den zu untersuchenden Gegenstandsbereichen und erforderte im Gegensatz zu T-MITOCAR noch die Begriffspaareinschätzung durch Probanden bezüglich der semantischen Ähnlichkeit bzw. Unähnlichkeit. Dies verlängerte den Auswertungsprozess erheblich. Zahlreiche Studien führten dabei zu homogenen, reliablen und validen Ergebnissen (vgl. Ifenthaler & Pirnay-Dummer, 2009). Das Instrument untersucht Eigenschaften sprachlicher Re-Repräsentationen, generiert dabei Gruppengraphen und vergleicht die Modellstrukturen von Gruppen innerhalb bestimmter Gegenstandsbereiche miteinander. Ifenthaler (2010) gibt einen Überblick an empirischen Studien, die die Anwendung der Graphentheorie fokussieren und verweist dabei auf die sehr geringe Anzahl solcher empirischer Untersuchungen im Bereich der Erziehungswissenschaft (Ifenthaler, 2010, S. 222). Die von Ifenthaler und Pirnay-Dummer zur Abgrenzung unterschiedlicher Expertisegrade (vgl. Pirnay-Dummer, 2006) sowie zur Überprüfung von Expertisezuwachs (vgl. Ifenthaler, 2006) entwickelten Kennwerte eigenen sich vor allem in Bezug auf lerntheoretische Fragestellungen (vgl. Ifenthaler, 2010). Auf Grundlage von Texten erstellt die Software assoziative Netzwerke, die sich bei

einer Textlänge von bereits 350 Wörtern als stabil erweisen (vgl. Pirnay-Dummer & Spector 2008). Die einzelnen automatisierten Analysevorgänge erfolgen in einzelnen Teilschritten (ebd., S. 11-24). Die Nomina werden aus dem Korpus herausgefiltert und bis auf den Wortstamm reduziert. Die daraus resultierende Begriffsliste wird auf die Häufigkeit hin analysiert. Die Anzahl der im Graphen erscheinenden Begriffe hängt von der Länge des Textes ab und umfasst nicht mehr als insgesamt 30 Wörter. Ein Algorithmus legt die Assoziiertheit dieser Begriffe fest. Mithilfe der einzelnen Teilschritte erfolgt eine Re-Repräsentation des Textes durch einen Graphen. Der Vergleich zweier Graphen erfolgt über sechs Kennwerte, die bestimmen wie ähnlich sich zwei Graphen auf semantischer sowie auf struktureller Ebene sind. Dabei bestimmen folgende Kennwerte die semantische Assoziiertheit: die Begriffsübereinstimmung (Concept Matching), die propositionale Übereinstimmung (Propositional Matching) sowie die ausbalancierte semantische Übereinstimmung (Balanced Semantic Matching). Die strukturelle Ähnlichkeit wird über folgende Kennwerte bestimmt: die Oberflächenstruktur (Surface Structure), die graphische Übereinstimmung (Graphical Matching), die Verdichtung der Knoten (Density of Vertices) und die strukturelle Übereinstimmung (Structural Matching) (vgl. Ifenthaler & Pirnay-Dummer, 2009). Im Folgenden erfolgt eine kurze Beschreibung der einzelnen Kennwerte.

4.1.1 Strukturelle Kennwerte

- Das *Surface Matching* vergleicht die Anzahl an Knoten zweier graphischer Re-Repräsentationen miteinander. Dabei berechnet sich die Surface Struktur (Oberflächenstruktur) aus der Summe aller Propositionen (vgl. Ifenthaler, 2006; Ifenthaler, 2008; Pirnay-Dummer & Ifenthaler, 2010). Der Ähnlichkeitsindex liegt zwischen 1 und 0 und ist ein Indiz für die Komplexität des Modells. Ein hoher Surface Matching Kennwert lässt auf eine ähnliche Komplexität zweier Modelle schließen.
- Das *Graphical Matching* vergleicht die Komplexität von Modellen strukturell miteinander. Dabei wird der kürzeste Pfad zwischen den am weitesten voneinander entfernt liegenden Knoten berechnet (ebd.). Das

Graphical Matching gibt einen Hinweis, für die Breite des konzeptuellen Wissens (vgl. Pirnay-Dummer & Ifenthaler, 2010, S. 101).

- Das *Structural Matching* vergleicht zwei Modelle in Bezug auf ihre interne Struktur strukturell miteinander (vgl. Pirnay-Dummer, 2012, S. 180). Es basiert auf der Ähnlichkeit nach Tversky (1977).
- Das *Gamma Matching* beschreibt die Dichte der Knoten innerhalb eines Modells. Es berechnet sich aus dem Quotienten der Eckpunkte innerhalb eines Graphen. Pirnay-Dummer (2006) zeigte, dass Expertenmodelle gewöhnlich einen Mittelwert von $s = 0.32$ haben (vgl. Pirnay-Dummer, 2006). Modelle mit einer Ähnlichkeit von $s = 1$ zeigen, dass beide Modelle dieselbe Dichte an Knoten haben. Eine Modellähnlichkeit von $s = 0$ zeigt, dass sich die Knotendichte beider Modelle völlig voneinander unterscheiden.

4.1.2 Semantische Kennwerte

- Das *Concept Matching* analysiert die konzeptuelle Verwendung zwischen zwei Modellen. Dies geschieht, indem die Anzahl übereinstimmender Konzepte mit den nicht übereinstimmenden Konzepten in Beziehung zueinander gebracht werden (vgl. Pirnay-Dummer, 2012).
- Das *Propositional Matching* vergleicht die propositionale Übereinstimmung zweier Modelle miteinander. Es berücksichtigt im Vergleich zu Concept Matching die Tversky-Ähnlichkeit (vgl. Pirnay-Dummer, 2012).
- Das *Balanced Semantic Matching* verbindet das Concept Matching und das Propositional Matching indem es sowohl Konzepte als auch Propositionen miteinander vergleicht (vgl. Pirnay-Dummer, 2012).

4.2 AKOVIA

Im Gegensatz zu T-MITOCAR eignet sich AKOVIA insbesondere bei größeren Datenmengen und führt die individuellen Ähnlichkeitswerte der externen Musterlösung gleich in eine Excel-Tabelle ein. Die Algorithmen sind identisch zu denen von T-MITOCAR (vgl. Pirnay-Dummer & Ifenthaler, 2010; Pirnay-Dummer, 2012). Der Benutzer lädt seine Daten über ein vorgefertigtes Muster auf

einen Server hoch und erhält per E-Mail eine Benachrichtigung wenn seine Auswertungen abholbereit auf dem Server bereitstehen. Dies stellt - im Gegensatz zu T-MITOCAR - eine zeitliche Verkürzung dar, da die einzelnen Ähnlichkeitskennwerte nicht mehr manuell in eine Exceldatei oder SPSS-Datei überführt werden müssen.

4.3 Qualitative Inhaltsanalyse

Die Forschungsperspektive, die auf einen bestimmten Gegenstandsbereich geworfen wird, entscheidet, welche Methoden der Datenerhebung und Interpretation herangezogen wird. Flick, Kardorff und Steinke (2010) geben hierfür einen Überblick in der sie u. a. die Qualitative Inhaltsanalyse, die Dokumentenanalyse und die objektive Hermeutik gegenüberstellen.

Die Qualitative Inhaltsanalyse eignet sich zur systematischen Analyse von Texten, indem diese durch zuvor festgelegte Regelwerke auf für die Fragestellung relevanten Informationen hin untersucht werden und dabei Kategorien gebildet werden. Mayring (2012) argumentiert, dass sich die Qualitative Inhaltsanalyse insbesondere eignet, im Rahmen von Mixed-Methods-Untersuchungen eingesetzt zu werden (ebd., S. 27). Folgende wesentlichen Bestandteile sind charakteristisch für die Qualitative Inhaltsanalyse: die Einordnung in ein Kommunikationsmodell, die Regelgeleitetheit, das Arbeiten mit Kategorien sowie die Berücksichtigung der Gütekriterien (vgl. Mayring 2008 a, S. 10, 2008b, 2012, S. 28 f). Das Einordnen in ein sprachtheoretisches Modell legt die Ziellegung der Analyse fest, d. h. es bestimmt welche Bestandteile für die schlussfolgernde Textanalyse wichtig sind (vgl. Mayring, 2012, S. 28). Die Vorgehensweise bei der Analyse mittels festgelegter Regeln umfasst die Bestimmung der Analyseeinheiten. Das Ablaufmodell umfasst folgende mögliche Herangehensweisen: die Zusammenfassung, die Explikation sowie die Strukturierung. Die zusammenfassende Analyse paraphrasiert, generalisiert sowie reduziert die in den Texten enthaltenen Information. Ziel dieser Analyse ist es, den Corpus so reduziert abzubilden, dass immer noch die wesentlichen Aspekte abgebildet sind (vgl. Mayring, 2008 b, S. 58). Die Explikation zieht zu unklaren Textausschnitten zusätzliche Informationen heran, um diese verständlicher zu machen. Die Strukturierung definiert die Kategorien, verwendet konkrete Ankerbeispiele aus

den gefundenen Textstellen und bestimmt die Kodierregeln. Es wird zwischen der formalen-, der inhaltlichen-, der typisierenden- sowie der skalierenden Strukturierung unterschieden.
Die Grounded Theory berücksichtigt im Gegensatz zur Qualitativen Inhaltsanalyse keine regelorientierte Analyse, sondern interpretiert die Textbestandteile auf Grundlage des zugrunde gelegten Materials (vgl. Bryant & Charmaz, 2007).

4.4 Methodologische Diskussion

Der folgende Teil ordnet die methodologischen Vorgehensweisen ein. Dabei müssen sich die methodologischen Zugänge immer an den gestellten Fragestellungen orientieren. Tashakoori & Teddlie (1998) differenzieren zwischen dem positivistischen Paradigma und dem konstruktivistischen Paradigma (ebd., 1998, S. 3). Erstes richtet sich rein quantitativ- und zweites rein qualitativ aus. Beide Ansätze verfolgen unterschiedliche Zielsetzungen bei der Generalisierbarkeit von Ergebnissen. Der quantitative Zugang fokussiert vor allem die Berücksichtigung der *externen Validität*. Hier soll sichergestellt werden, dass eine möglichst gültige Schlussfolgerung von der gezogenen Stichprobe auf die Zielpopulation gezogen werden kann. Der qualitative Zugang beabsichtigt das Einhalten der *internen Validität*. Diese bemüht sich um möglichst gültige Aussagen in Bezug auf den Einzelfall (vgl. Bortz & Döring, 2003). Eine Methodenkombination (welche qualitative sowie quantitative Zugänge integriert) ermöglicht einen potenziellen Mehrgewinn an wissenschaftlichen Erkenntnissen, welche auf der Grundlage eines rein quantitativen- oder rein qualitativen Zugangs nicht möglich gewesen wären (vgl. Teddlie & Tashakkorie, 2010). Greene, Caracelli & Graham (1989) unterscheiden zwischen fünf Zielsetzungen des Mixed Method-Ansatzes. Erstens, um mittels Triangulation konvergente Ergebnisse zu erzielen. Zweitens, um mittels der Mischung quantitativer sowie qualitativer Zugänge unterschiedliche Facetten hinsichtlich des Untersuchungsgegenstandes zu diskutieren. Drittens, um widersprüchliche Befunde einzuordnen. Viertens, die Innovation von Forschungsmethoden voranzutreiben. Fünftens, um einen breiteren Zugang des Gegenstandsbereiches zu erzielen (vgl.Tashakkori &Teddlie, 1998, S. 43). Vorliegende Arbeit fokussiert dabei letzteres.

Im Bereich der Leistungsdiagnostik würde es ein quantitativer Zugang ermöglichen, auf Grundlage einer sehr großen Stichprobe systematische Aussagen zu treffen. Ein qualitativer Zugang würde es ermöglichen, auf Grundlage von Einzelfallanalysen detailliertere Erkenntnisse zu erfassen, die auf rein quantitativer Ebene nicht möglich wären. Zum Beispiel einer detaillierteren Analyse der zugrunde gelegten Bewertungskriterien oder eine differenziertere Untersuchung der diagnostischen Expertise der Lehrkraft. Eine Metthodenkombination ermöglicht tiefere Erkenntnisse, um differenziertere Aussagen zu treffen.

Gläser-Zikuda (2008) verweist auf die Dringlichkeit im Bereich der Lehr-Lernforschung qualitative Zugänge zu wählen, da dies neue Möglichkeiten eröffnet insbesondere in Untersuchungsdesigns, welche Einzelfallanalysen oder Feldforschung fokussieren, da dies neue Möglichkeiten eröffnet, neue Erkenntnisse, insbesondere in Untersuchungsdesigns, welche Einzelfallanalysen oder Feldforschung fokussieren, zu erzielen. Bislang spezialisierte sich die Lehr-Lernforschung stark an quantitativen Zugängen (ebd, S. 65; Helmke & Weinert, 1997). Vertreter, die sich in der Lehr-Lernforschung an qualitativen Zugängen orientierten sind beispielweise Wild (2001) und Prenzel et al., (2002). Flick, Kardorff & Steinke (2010) verorten die unterschiedlichen qualitativen Zugänge in den theoretischen Rahmen.

Gläser-Zikuda, Seidel, Rohlfs, Gröschner & Ziegelbauer (2012) verdeutlichen drei Trends, welche bei Mixed-Methods Studien im empirischen Bildungskontext erkennbar sind: die Etablierung qualitativer Forschungsmethoden, die stärkere Wahrnehmung quantitativer Forschungsmethoden, sowie die Kombination qualitativer und quantitativer Verfahren. Vorliegende Arbeit fokussiert letzteres.

5 Fragestellungen und Hypothesen

Ausgehend von den theoretischen Vorüberlegungen, die die vorliegende Forschungsarbeit zunächst in den konstruktivistischen Lernansatz auf Grundlage der mentalen Modellbildung und des selbstregulierten Lernens einbettet, sowie in die Wissensrepräsentation welche für die Feststellung von Leistungen notwendig ist, wurde die diagnostische Expertise als notwendige Voraussetzung von Lehrkräften diskutiert. Der empirische Forschungsstand zeigte, dass die lehrerbasierte Leistungsdiagnostik den wissenschaftlichen Standards nicht genügt. Der aktuelle Forschungsstand hinsichtlich einer möglichen technologiegestützten Leistungsdiagnostik zeigte eine Rarität an möglichen Auswahlinstrumenten. Auf Grundlage der theoretischen Vorüberlegungen ergibt sich die Fragestellung, inwiefern sich die Leistungsfeststellung schriftlicher Prüfungsleistungen durch im Rahmen der mentalen Modelle entwickelten Technologien unterstützen lässt. Die Fragestellungen der empirischen Studien im Hochschul- sowie im Schulkontext verfolgen dabei unterschiedliche Fragestellungen. Zunächst folgte an der Hochschule eine erste Überprüfung, ob sich die Dozentenbasierte Leistungsfeststellung überhaupt technologiegestützt unterstützen lässt.

Im **Hochschulkontext** stehen folgende Fragen im Vordergrund.

- Lässt sich die (inhaltlich orientierte) Leistungsfeststellung von Hochschuldozenten durch T-MITOCAR abbilden?
- Lassen sich die Kriterien durch eine schriftliche Musterlösung abbilden? (explorative Fragestellung)

Daraus leiten sich folgende Kernhypothesen ab:

- H_1: Es besteht ein Zusammenhang (mithilfe der Regressionsanalyse und des korrigierten R^2) zwischen den inhaltlich orientierten Bewertungen (die durch die Hochschuldozenten ermittelt wurden) und den durch T-MITOCAR ermittelten semantischen Kennwerten.
- H_{01}: Es besteht kein Zusammenhang (mithilfe der Regressionsanalyse und des korrigierten R^2) zwischen den inhaltlich orientierten Bewertungen (die durch die Hochschuldozenten ermittelt wurden) und den durch T-MITOCAR ermittelten semantischen Kennwerten.

Nach dieser ersten Überprüfung verfolgen die Studien im **Schulkontext** die Frage:

- Lässt sich die Leistungsfeststellung von Lehrkräften durch T-MITOCAR abbilden?

Dabei ergeben sich verschiedene Kernyhpothesen, je nachdem, ob die Lehrkräfte in ihren Kriterien eine strikte Trennung von inhaltlichen und strukturellen Bewertungen vornehmen oder nicht. Erfolgt eine strikte Trennung ergeben sich daraus folgende Hypothesen:

- H_2: Es besteht ein Zusammenhang (mithilfe der Regressionsanalyse und des korrigierten R^2) zwischen den inhaltlich orientierten Bewertungen (die durch die Lehrkräfte ermittelt wurden) und den durch T-MITOCAR ermittelten semantischen Kennwerten.
- H_{02}: Es besteht kein Zusammenhang (mithilfe der Regressionsanalyse und des korrigierten R^2) zwischen den inhaltlich orientierten Bewertungen (die durch die Lehrkräfte ermittelt wurden) und den durch T-MITOCAR ermittelten semantischen Kennwerten.

- H_3: Es besteht ein Zusammenhang (mithilfe der Regressionsanalyse und des korrigierten R^2) zwischen den strukturell orientierten Bewertungen (die durch die Lehrkräfte ermittelt wurden) und den durch T-MITOCAR ermittelten strukturellen Kennwerten.
- H_{03}: Es besteht kein Zusammenhang (mithilfe der Regressionsanalyse und des korrigierten R^2) zwischen den strukturell orientierten Bewertungen (die durch die Lehrkräfte ermittelt wurden) und den durch T-MITOCAR ermittelten strukturellen Kennwerten.

Erfolgt keine stringente Trennung nach inhaltlichen und strukturellen Kriterien, ergeben sich folgende Hypothesen:

- H_4: Es besteht ein Zusammenhang (mithilfe der Regressionsanalyse und des korrigierten R^2) zwischen den Gesamtbewertungen (die durch die Lehrkräfte ermittelt wurden) und den durch T-MITOCAR ermittelten semantischen Kennwerten.
- H_{04}: Es besteht kein Zusammenhang (mithilfe der Regressionsanalyse und des korrigierten R^2) zwischen den Gesamtbewertungen (die durch die

Lehrkräfte ermittelt wurden) und den durch T-MITOCAR ermittelten semantischen Kennwerten.

- H_5: Es besteht ein Zusammenhang (mithilfe der Regressionsanalyse und des korrigierten R^2) zwischen den Gesamtbewertungen (die durch die Lehrkräfte ermittelt wurden) und den durch T-MITOCAR ermittelten strukturellen Kennwerten.
- H_{05}: Es besteht kein Zusammenhang (mithilfe der Regressionsanalyse und des korrigierten R^2) zwischen den Gesamtbewertungen (die durch die Lehrkräfte ermittelt wurden) und den durch T-MITOCAR ermittelten strukturellen Kennwerten.

Im Zentrum der qualitativen Zugänge der Schulstudien standen folgende Fragestellungen:

- Welche Kriterien legen Lehrkräfte bei der Bewertung textbasierter Schülerleistungen in unterschiedlichen Unterrichtsfächern an?
- Wie gehen Lehrkräfte vor, von der Erstellung der Klausurfragen bis hin zu der Bewertung und Benotung eben dieser?
- Welche Bewertungsqualität haben Lehrkräfte in unterschiedlichen Unterrichtsfächern?

5.1 Pädagogische Intervention

Nachdem eine Überprüfung der eben dargestellten Fragestellungen erfolgte, interessierte im Schulkontext zudem die Frage:

- Können die Bewertungskriterien von Lehrkräften durch eine gezielte Intervention verändert werden?

Damit stand die Veränderung der durch die Lehrkräfte erstellten Bewertungskriterien sowie der textbasierten Musterlösung im Vordergrund.
Daraus leiten sich folgende Kernhypothese ab:

- $H_{6:}$ Es besteht nach der Intervention ein höherer Zusammenhang (mithilfe der Regressionsanalyse und des korrigierten R^2) zwischen den Bewertungen (die durch die Lehrkräfte ermittelt wurden) und den durch T-MITOCAR ermittelten Kennwerten.

- H_{06}: Es besteht nach der Intervention kein höherer Zusammenhang (mithilfe der Regressionsanalyse und des korrigierten R^2) zwischen den Bewertungen (die durch die Lehrkräfte ermittelt wurden) und den durch T-MITOCAR ermittelten Kennwerten.

- $H_{7:}$ Es besteht nach der Intervention ein höherer Zusammenhang (mithilfe der Regressionsanalyse und des korrigierten R^2) zwischen den Bewertungen (die durch die Lehrkräfte ermittelt wurden) und den durch T-MITOCAR ermittelten Kennwerten auf Grundlage der nach der Intervention veränderten Musterlösung als mit der vor der Intervention erstellten Musterlösung
- H_{07}: Es besteht nach der Intervention kein höherer Zusammenhang (mithilfe der Regressionsanalyse und des korrigierten R^2) zwischen den Bewertungen (die durch die Lehrkräfte ermittelt wurden) und den durch T-MITOCAR ermittelten Kennwerten auf Grundlage der nach der Intervention veränderten Musterlösung als mit der vor der Intervention erstellten Musterlösung.

Interventionsmaßnahmen im pädagogisch-psychologischen Kontext haben die Zielsetzung, Fähigkeiten zu fördern, mögliche Defizite vorzubeugen bzw. bereits vorhandene Defizite zu eliminieren, sowie Fähigkeiten wieder aufzubauen (vgl. Hager & Hasselhorn, 2008). Interventionen lassen sich in direkte und indirekte Interventionen unterscheiden (vgl. Schmidt & Otto, 2010). Direkte Interventionen fokussieren die Schülerinnen und Schüler, wohingegen die indirekten Interventionen bei den Lehrkräften ansetzen. Hagenauer (2010) differenziert zwischen Kurzzeit- und Langzeitinterventionen. Sie definiert dabei eine Kurzzeitstudie als eine Prä-Post-Untersuchung, wohingegen sie einer Langzeitstudie noch eine Follow-Up-Messung hinzufügt. Interventionen können auf unterschiedlichen Ebenen stattfinden: der Mikroebene, der Mesoebene sowie der Makroebene (vgl. Leutner, 2010, S. 63; Fend, 2008). Die Mikroebene nimmt den Lernenden in den Fokus, die Mesoebene fokussiert die Institution Schule, wohingegen die Makroebene das Schulsystem in den Fokus rückt. Trittel (2010) verweist auf den Nutzen sowie die Problematik von Einzelfallanalysen. Sie

begründet diese Vorgehensweise wenn ein explorativer Charakter vorliegt (ebd., S. 281).

In der vorliegenden Forschungsarbeit im schulischen Kontext liegt eine indirekte Intervention vor, da zwar die Lehrkräfte sowie deren Bewertungsqualität im Prozess der Leistungsbewertung in den Blick genommen werden, die Schüler dadurch jedoch eine möglicherweise optimierte Bewertung ihrer Lernergebnisse erfahren. Die Schulungsmaßnahme intendiert demnach eine Steigerung der Bewertungsqualität von Seiten der Lehrkräfte. Es liegt im Schulkontext eine Kurzzeitintervention vor. Die Maßnahme hat im Laufe des Schuljahres nur einmal stattgefunden und es wurde in einem Pre-Post-Design die Bewertungsqualität sowie die eigentliche Bewertung der Lehrkräfte erfasst. Die Untersuchung befindet sich auf der Mesoebene, da sie auf Schulebene die Bewertungsqualität der Lehrkräfte fokussiert. Es liegen Einzelfallanalysen vor, da Lehrkräfte nicht systematisch erfasst werden.

Engel & Hurrelmann (1989) unterscheiden zwischen vier Arten von Interventionen. Diese unterteilen sie zunächst in die Zieldimension: personale und soziale Ressourcen sowie in die Art der Maßnahme: präventiv oder korrektiv. Vorliegende Arbeit hat die Zielsetzung die konkrete Bewertung von Lehrkräften (konkret deren textbasierten Musterlösungen sowie deren Bewertungskriterien) gezielt zu verändern. Diese fokussiert nach Engel & Hurrelmann (1989) das Training individueller Kompetenzen (vgl. Schwarzer & Buchwald, 2006, S. 581).

Hascher (2010) differenziert zwischen forschungs- versus praxisorientierten Ansätzen von Interventionsmaßnahmen. Diese unterscheidet sie anhand von fünf Merkmalen. Vorliegende Intervention basiert auf theoretischen Vorannahmen und lässt sich demnach einer forschungsorientierten Maßnahme zuordnen.

6 Methode

Das folgende Kapitel beschreibt die empirische Überprüfung im Hochschulkontext sowie im Schulkontext. Diese fokussiert die konkrete Bewertungssituation in unterschiedlichen Gegenstandsbereichen.

Das Untersuchungsdesign der vorliegenden Forschungsarbeit bestand aus zwei Teilstudien im Hochschulkontext (N = 228 Studierende) sowie aus vier Teilstudien im Schulkontext (N = 105 Schüler). Dabei standen im Schulkontext sechs Einzelfallanalysen im Vordergrund in denen die Bewertungsqualität von Lehrkräften sowie deren praktische Umsetzung untersucht wurden. An der Hochschule erfolgte zunächst ein quantitativ-orientierter Zugang um die in Kapitel fünf gestellten Fragen zu beantworten. Die Schulstudien beinhalten sowohl quantitative- als auch qualitative Zugänge. Dies war für die dort gestellten Fragestellungen (vgl. Kapitel fünf) notwendig.

Zunächst erfolgt die Darstellung der empirischen Überprüfung im Hochschulkontext und daran anschließend im Schulkontext. Beide umfassen jeweils die Durchführung und Darstellung der empirischen Befunde sowie einer anschließenden Diskussion und Einbettung in den theoretischen Hintergrund.

Diese erfolgen im Hochschul- und Schulkontext jeweils getrennt voneinander.

6.1 Konzeptionen der Hochschuluntersuchungen

Die Konzeptionen der beiden durchgeführten Studien im Hochschulkontext umfassen jeweils die konkrete Bewertungssituation (Aufgabenstellung, Bewertungskriterien), das Untersuchungsdesign, die Stichprobenbeschreibungen sowie das methodische Vorgehen. Letzteres schließt die Datenerhebung (6.1.5.1), die Validitätsbestimmung der Musterlösungen (6.1.5.2), die methodenkritischen Anmerkungen (6.1.5.3) sowie die Auswertungsverfahren (6.1.5.4) ein. Zunächst erfolgt die Darstellung der Konzeption der ersten Studie im Hochschulkontext und daran anschließend die der zweiten Studie.

6.1.1 Konkrete Bewertungssituation der ersten Untersuchung

Um eine konkrete Bewertungssituation für Lehrkräfte im Hochschulkontext zu ermöglichen, wurde eine klassische Prüfungssituation in Form einer Klausur am Ende des Semesters zur Untersuchung herangezogen. Diese erfasste die im Laufe des Semesters behandelten Themen der Einführungsvorlesung „Grundlagen der Schulpädagogik“ in den Studiengängen Lehramt an Regelschulen und Gymnasien, welche im Wintersemester 2008/ 09 an der Friedrich-Schiller-Universität Jena stattfanden. Die Themen der Vorlesung umfassten unter anderem: Schultheorien, diagnostische Verfahren, Unterrichtsmethoden, didaktische Modelle, selbstreguliertes Lernen und Metakognition sowie Schul- und Unterrichtsforschung. Begleitend zur Vorlesung nahmen die Lernenden wahlweise an begleitenden Tutorien teil. Zudem bearbeiteten die Studierenden im Laufe des Semesters Aufgabenstellungen, in denen sie in Form von Texten ihr Wissen externalisierten. Am Ende des Vorlesungszeitraums schrieben sie eine Abschlussklausur. Die Klausur beinhaltete unter anderem zwei textproduzierende Aufgabenformate, in denen die Probanden ihr deklaratives Wissen aus den Bereichen des selbstregulierten Lernens sowie der Metakognition in Textform abbildeten. Diese hatten sie in Vorbereitung auf die Klausur mit vorgegebener Literatur bearbeitet. Des Weiteren umfasste die Klausur Aufgaben, in denen die Studierenden ihr Wissen in Form von offenen sowie geschlossenen Antwortformaten darstellten. Für die vorliegende Untersuchung wurden die

textproduzierenden Aufgabenformate herangezogen. Dabei gab es folgende Problemstellungen:

> *Bitte bearbeiten Sie (auf jeweils 2-3 Seiten) die folgenden Aufgabenstellungen in Textform.*
>
> *1. Selbstreguliertes Lernen (10 Punkte)*
> *Erläutern Sie die einzelnen Phasen des selbstgesteuerten Lernens. Gehen Sie außerdem auf die Probleme und Grenzen der Selbstregulation und des Selbstregulationsmodells ein.*
>
> *2. Metakognition (10 Punkte)*
> *Erläutern Sie die wesentlichen und charakteristischen Merkmale metakognitiven Lernens.*

Zwei Hochschuldozenten bewerteten die in Form von Texten bearbeiteten Aufgabenstellungen unabhängig voneinander mithilfe von ihnen vorgegebenen inhaltlich orientierten Kriterienkatalogen (vgl. Tabelle 1 und 2). Bei der ersten Aufgabenstellung waren die Kriterien zunächst in eine Definition sowie in die einzelnen drei Phasen des selbstregulierten Lernens unterteilt. Daran anschließend erfolgten inhaltliche Punkte für Probleme und Grenzen des Selbstregulationsmodells. Die zweite Aufgabenstellung unterteilte sich ebenso in einzelne Bereiche - beispielsweise der Definition, metakognitives Wissen, prozedurales Wissen, Selbstwirksamkeit, Selbstmanagement usw. - welche wiederum jeweils einzelne Aspekte umfassten - die von den Studierenden in ihren Texten abgebildet sein sollten.

Nachdem die Hochschuldozenten auf Grundlage der Kriterienkataloge den schriftlich basierten Prüfungsergebnissen Punkte vergaben, wurden diese später mit den durch T-MITOCAR identifizierten Kennwerten verglichen. Als Außenkriterium um die Ähnlichkeit der einzelnen Prüfungsleistungen zu bestimmen wurden textbasierte Musterlösungen, sowie die jeweiligen Lehrtexte, die den Studierenden zur Vorbereitung dienten, herangezogen. Die Beschreibung der Bestimmung der inhaltlichen Vollständigkeit sowie Richtigkeit der einzelnen Musterlösung erfolgt in Abschnitt 6.1.5.2.

Tabelle 1 Bewertungskriterien des selbstreguliertes Lernens

		Punkte
Definition	Selbstreguliertes Lernen stellt einen ständigen „Soll-Ist-Vergleich“ dar zwischen dem aktuellem (Lern-) Stand sowie dem Zielzustand. Alternativen: - adaptive Zielverfolgung/ prozessualer Charakter - Der Lerner überwacht den Fortgang seiner Lernprozesse selbst - Kognitive, motivationale sowie emotionale Faktoren beeinflussen das selbstgesteuerte Lernen	0.5
	1) Präaktionale Phase	0.5
Zusatzpunkte	Emotionen, die sich bei schwierigen Aufgaben als Angst/ Unlust auswirken und bei herausfordernden/ interessanten Aufgaben als Neugier/ Hoffnung auf Erfolg/ (Vor-) Freude	0.5
	Motivation beeinflusst die Anstrengungsbereitschaft (im Zusammenhang mit der Zielsetzung) Alternative: - Selbstwirksamkeit	0.5
	Lernziele formulieren/ festlegen (= Kernstück der Selbstregulation) (1 P.)	1.0
	Planung (1 P.)	1.0
	2) Aktionale Phase	0.5
	Einsatz (aufgabenspezifischer) Lernstrategien (1 P.) -kognitive, metakognitive, ressourcenbezogene Strategien (intern, extern)	1.0
	Self-Monitoring/ Überwachung der Lernhandlungen	1.0
	3) Post-Aktionale Phase	0.5
	Selbstreflexion (der erzielten Ergebnisse) Alternative: - (Soll-Ist-Vergleich)	1.0
	Ziehen von Konsequenzen im Hinblick auf weitere Lernprozesse	1.0
Probleme der Selbstregulation / Probleme im Zusammenhang mit dem Selbstregula-tionsmodell	Erfolg selbstregulierten Lernens hängt ab von hoher Zielbildung Alternativen: - Genaues Monitoring - Effektive Regulation - Es gibt keine Theorie der Selbstregulation, sondern verschiedene Ansätze verschiedener Autoren - Nur teilweise Überlappung der einzelnen Ansätze - Verwendung des Begriffs Selbstregulation und Bezug auf Teilkonstrukte davon (wie z. B. Lernstrategien, Selbstkonzept, Interesse) Mindestens 2 Punkte müssen angesprochen sein	
	Summe + Bonus	10.0 + 1.0

Tabelle 2 Bewertungskriterien der Metakognition

		Punkte
Definition	Flavell/ Brown (1978) Die Fähigkeit das eigene Lernen zu beobachten und zu evaluieren sowie Lösungswege zu entwerfen.	1.0
Metakognitives Wissen	*Metakognitives Wissen = Wissen über das eigene Denken/ das anderer Personen = Wissen über Anforderungen/ eigene Kognition	1.0
Prozedurales Wissen	*Prozedurales (metakognitive) Wissen/ Strategien = Kontrolle bei der Bearbeitung von Aufgaben = Regulierung bei der Bearbeitung von Aufgaben	1.0
Selbstwirksamkeit	*Selbst-Wirksamkeit = „was weiß ich eigentlich“ = “wie denke ich“ = “wann wende ich Wissen/ Strategien an?“ = “warum wende ich Wissen/ Strategien an?“	1.0
Selbstmanagement	Selbst-Management = effektive Organisation kognitiver/ metakognitiver Prozesse	1.0
Selbstbewertung	Selbst-Bewertung = Bewusstwerden motivationaler, emotionaler Zustände bei der Bearbeitung einer Aufgabe	1.0
Lernstrategien	*Lernstrategien: Wiederholungsstrategien, Elaborationsstrategien, Organisationsstrategien, Kontrollstrategien	1.0
Metakognitive Strategien	*Metakognitive Strategien: Setzen von Zielen, Eigenständige Organisieren von Informationen, Selbstbeobachtung, Selbstbeurteilung/ Reflexion	1.0
Strategische Lerner	*Strategische Lerner: Wissen über sich selbst, Wissen über unterschiedliche Typen/ Aufgaben, Wissen über (Lern-)Strategien, Vorwissen, Wissen über jetzige und zukünftige Situationen	1.0
Strategische Wissensnutzer	Strategische Wissensnutzer: Lernexperte/ Lerner nutzt sein Wissen strategisch	1.0
Zusatz	Selbstreguliertes Lernen: Lernexperte/ Lerner lernt selbstreguliert Lernexperte/ erfolgreicher Lerner lernt zielorientiert Suchen nach Informationen zur Lösung einer Aufgabe Nutzen viele strategische Verhaltensweisen zur Optimierung von Lernerfolgen	0.5
	Metakognitives Lernen: Erwerb von (Lern-) Strategien zur Verarbeitung von Informationen (0.5 P.) **oder** Erleichterung von Problemlösen (0.5 P.)	0.5
	Metakognitives Wissen: Wissen über das eigene Wissen und Verhalten (0.5 P.) Metakognitive Fertigkeiten: Handeln in Bezug auf eine Aufgabenstellung (0.5 P.) Metakognitive Erfahrungen: kognitive/ emotionale Beurteilung der (augenblicklichen) Situation (0.5 P.)	0.5
	Summe + Bonus	10,0 + 0,5

*mind. 2 Punkte müssen angesprochen sein

6.1.2 Konkrete Bewertungssituation der zweiten Untersuchung

Der folgende Abschnitt beschreibt die konkrete Bewertungssituation der zweiten Studie im Hochschulkontext. Hier wurde wie in der ersten Untersuchung eine klassische Prüfungssituation gewählt. Diese fokussierte die Abschlussklausur der Vorlesung „Pädagogische Psychologie“ in den Studiengängen Lehramt an Regelschulen und Gymnasien, welche im Sommersemester 2009 an der FSU Jena stattfand. Die Klausurinhalte erfassten die behandelten Themen der Vorlesung. Vorlesungsgegenstand waren unter anderem zentrale Themen der Pädagogischen Diagnostik, die Motivation, Forschungsparadigmen, das selbstregulierte Lernen und Lernstrategien. Am Ende des Vorlesungszeitraums schrieben die Studierenden dann eine Abschlussklausur. Die Klausur beinhaltete unter anderem zwei textproduzierende Aufgabenformate, in denen die Studierenden wahlweise ihr deklaratives Wissen entweder im Bereich des selbstregulierten Lernens oder im Bereich der Lernstrategien in Textform externalisierten. Beide Inhaltsbereiche wurden mit entsprechend vorgegebener Literatur vorbereitet. Für die vorliegende Untersuchung wurden die textproduzierenden Aufgabenformate herangezogen.
Die Aufgabenstellungen waren die folgenden:

> *Bitte bearbeiten Sie wahlweise eine der folgenden Aufgabenstellungen.*
>
> 1. *Erläutern Sie die einzelnen Phasen des selbstgesteuerten Lernens (siehe Schmitz & Schmidt 2007, S. 9-19). Gehen Sie außerdem auf die Probleme und Grenzen der Selbstregulation und des Selbstregulationsmodells ein (Schmitz, Landmann & Perels 2007, S. 312-326). (10 Punkte)*
> 2. *Erläutern Sie alle nach Mandl & Friedrich (2006, S. 1-23) unterschiedenen Lernstrategien. (10 Punkte)*

Auch diese wurden von jeweils zwei Hochschuldozenten mithilfe von vorgegebenen inhaltlich orientierten Kriterienkatalogen bewertet. Sie sind in den Tabellen 3 und 4 abgebildet. Im Vergleich zur ersten Untersuchung wurden die Kriterien des selbstregulierten Lernens auf Grundlage der Rückmeldungen der Bewerter angepasst. Da die Definition nicht direkt in der Aufgabenstellung verlangt war, wurde diese nicht weiter im Kriterienkatalog angegeben. Außerdem wurden die einzelnen inhaltlich orientierten Aspekte weiter ausdifferenziert. Der zweite Kriterienkatalog (vgl. Tabelle 4) umfasst die einzelnen Lernstrategien und ebenso zu jeder Lernstrategie einzelne inhaltliche Aspekte. Diese sollten von den Studierenden in Bezug auf die einzelnen Strategien in ihren textbasierten

Prüfungsleistungen enthalten und richtig angesprochen sein. Im Vergleich zur vorherigen Studie wurden die Aufgabenstellung Metakognition durch Lernstrategien ersetzt, da sich bei der Bearbeitung Schwierigkeiten ergeben hatten.

Tabelle 3 Bewertungskriterien des selbstreguliertes Lernens

		Punkte
Präaktionale Phase	**1) Präaktionale Phase** (1 P. x 0.5) Vorbereitungsphase/ Planung (1 P. x 0.5) Emotionen/ Motivation (1 P. x 0.5) Einschätzung der Selbstwirksamkeit (1 P. x 0.5) Ressourcenüberprüfung: (1 P. x 0.25) - Einschätzung notwendiger Strategien - Einschätzung des Vorwissens Einschätzung: (1 P. x 0.5) - der Aufgabe (schwierig/ leicht) - der Zeit, die benötigt wird - der Anstrengungsbereitschaft - Zielformulierung	2.75
Aktionale Phase	**1) Aktionale Phase** (1 P. x 0.5) Aufgabenbearbeitung (1 P. x 0.5) Anwendung aufgabenspezifischer Lernstrategien (1 P. x 0.5) Self-Monitoring/ Überwachen der Lernhandlung/ (1 P. x 0.5) Beobachten oder Aufzeichnen des eigenen Lernens. (1P. x 0.5) Vergleich des Istzustandes mit dem Sollzustand (Zusatz) Selbstreflexive Prozesse: Attributionen/ Ursachenbeschreibung (Zusatz)	2.5
Post-Aktionale Phase	**Post-Aktionale Phase** (1 P. x 0.5) Reflexion (1 P. x 0.75) Ziehen von Konsequenzen auf weiteres Lernen/ Lernprozesse (1 P. x 0.75) Evaluation der Lernprozesse (1 P. x 0.75) - Subjektive Einschätzung (Lernzufriedenheit) - Quantitative Leistung (Menge des Gelernten) - Qualitative Leistung (Ausmaß des Verstehens) Vergleich des Istzustandes mit dem Sollzustand (Zusatz) Selbstreflexive Prozesse: Attributionen/ Ursachenzuschreibung (Zusatz)	2.5
Probleme der Selbstregulation	**Probleme der Selbstregulation/ Probleme im Zusammenhang mit dem Selbstregulationsmodells** Nicht genügend Zielbildung (1 P. x 0.75) Mangelndes Self-Monitoring (1 P. x 0.75) Ineffektive Regulation (1 P. x 0.75) Schlechte Zielwahl (Zusatz) Es gibt keine Theorie der Selbstregulation, sondern lediglich verschiedene Ansätze verschiedener Autoren (1 P. x 0.75) Nur teilweise Überlappung der einzelnen Ansätze (1 P. x 0.75) Verwendung des Begriffs „Selbstregulation“ und lediglich Bezug auf Teilkonstrukte davon (wie z. B. Lernstrategien, Selbstkonzept, Interesse) (Zusatz)	4.5
	Summe	12.25

Tabelle 4 Bewertungskriterien der Lernstrategien

		Punkte
Kognitive Lernstrategien	Elaboration (1 P. x 0.4) Vorwissen aktivieren (1 P. x 0.4) Fragenstellen/ Notizenmachen (1 P. x 0.4) Vorstellungsbilder/ Imagery-Strategien/ Mnemotechniken (1 P. x 0.4) Wiederholungsstrategien/ Auswendiglernen (1 P. x 0.4)	2.
	Organisation/ Strukturierung (1 P. x 0.5) Wissen organisieren/ strukturieren (1 P. x 0.5) z. B. Zusammenfassen v. Texten (1 P. x 0.5) Nutzung v. Schemata (1 P. x 0.5) Strategien der externen Visualisierung (Mind Maps/ Concept Maps (1 P x 0:5)	2.5
	Selbstkontroll-/ Selbstregulation (1 P. x 0.5) Planung (1 P. x 0.5) Überwachung (1 P. x 0.5) Bewertung (1 P. x 0.5) Regulation (1 P. x 0.5)	2.5
	Wissensnutzung (1 P. x 0.5) Anwendung und Transfer von Wissen (1 P. x 0.5) Transferangemessene Verarbeitungsstrategien: - Das Lösen von Problem (1 P. x 0.5) - Das Schreiben von Texten (1 P. x 0.5) Das Argumentieren/ Diskutieren im sozialen Kontext (1 P. x 0.5)	2.5
Motivation, Emotion (1 P. x 0.5)	Individuelle Motivation: (1 P. x 0.5) - Intrinsische/ extrinsische Motivation - Thematisches Interesse - Ziele/ Bedürfnisse Motivationale Charakteristika der jeweiligen Lernumgebung (1 P. x 0.5) Motivational-emotionale Bedingungen beeinflussen: (1 P. x 0.5) - Die Anstrengung/ Ausdauer der Aufgabenwahl (Schwierigkeit, Inhalt der von einer Person gewählten Aufgabe) indirekt - Die Wahl entsprechender kognitiver und metakognitiver Lernstrategien	2
Kooperatives Lernen (1 P. x 0.5)	Kooperative Lernformen wirken auf Motivation & Kognition (1 P. x 0:5) Sozial-interaktive Lernformen wirken sich positiv aus auf: (1 P. x 0.5) - Die Motivation, selbst zu lernen - Die Motivation, andere zum Lernen zu motivieren - Die Motivation, anderen bei Lernen zu helfen Auswirkungen kooperativen Lernens: (1 P. x 0.5) - Das individuelle Generieren von Elaborationen - Das gegenseitige Erklären (*peer tutoring*) Das Lernen am Modell (*peer modeling*)	2
Nutzung von Ressourcen (1 P. x 0.5)	Zeit (1 P. x 0.5) Externe Speicher in Form von Notizen/ Datenbanken Nutzung bestimmter „tools“ wie beispielsweise PC (1 P. x 0.5) Management digitaler Lernressourcen (1 P. x 0.5) Lernumgebung als Ressource für Lernen (1 P. x 0.5)	2
	Summe + Bonus	15.5

6.1.3 Untersuchungsdesign

Das Untersuchungsdesign (vgl. Abbildung 1) beinhaltet in der ersten Hochschulstudie ein *within-subject design*. Das heißt dieselben (N = 45) Probanden externalisierten ihr Wissen zu beiden Aufgabenstellungen (selbstreguliertes Lernen sowie Metakognition). Die zweite Studie im Hochschulkontext umfasste ein *between-subject-design.* Unterschiedliche Probanden bildeten ihr Wissen wahlweise entweder bezüglich des selbstregulierten Lernens oder bezüglich der Lernstrategien ab.

Abbildung 1 Untersuchungsdesign

	Selbstreguliertes Lernen	**Metakognition**	**Lernstrategien**
Studie 1	N = 45	N = 45	-
Studie 2	N = 103	-	N = 70

6.1.4 Stichprobenbeschreibungen

6.1.4.1 Stichprobe der ersten Untersuchung im Hochschulkontext

Die Stichprobe umfasst 45 Studierende (33 weibliche, 12 männliche), die die Vorlesung „Grundlagen der Schulpädagogik“ besuchen und hier einen Leistungsschein erwerben. Das Durchschnittsalter der Studierenden beträgt ca. 22 Jahre (Min = 19; Max = 29, SD = 2.25).

6.1.4.2 Stichprobe der zweiten Untersuchung im Hochschulkontext

Die zweite Studie umfasst 183 Studierende, die die Vorlesung „Pädagogische Psychologie“ besuchen und hier einen Leistungsschein erwerben. Das Durchschnittsalter der Studierenden liegt bei ca. 21 Jahren (Min = 18; Max = 32, SD = 1.84). Davon sind N = 90 männlich und N = 92 weiblich. Eine Person gab das Geschlecht nicht an.

6.1.5 Methodisches Vorgehen

6.1.5.1 Datenerhebung

Die Studierenden bearbeiteten die Aufgabenstellung in einer für sie gewöhnlichen Prüfungssituation handschriftlich. Dies sollte ungewohnte Nebeneffekte verhindern. Die Klausuren wurden digitalisiert und anonymisiert.

6.1.5.2 Validitätsbestimmung der Musterlösung

Zur Validitätsbestimmung der textbasierten Musterlösung wurde diese von einem Experten aus dem gesuchten Gegenstandsbereich auf die inhaltliche Richtigkeit und Vollständigkeit hin geprüft und eingeschätzt. Dies erfolgte, indem die Musterlösung daraufhin eingeschätzt wurde, inwiefern alle in den Kriterien enthaltenen Aspekte auch im Außenkriterium enthalten waren. Zur Einschätzung wurde die Kategorie „trifft zu“ bzw. „trifft nicht zu“ gewählt. Zudem schätzte der Experte die Kriterien auf Grundlage des Lehrtextes daraufhin ein, ob die wesentlichen Aspekte in den Kriterien abgebildet waren. Auf Basis dieser Einschätzungen wurde die Musterlösung durch den Experten vorgenommen.

6.1.5.3 Methodenkritische Anmerkungen

Dieser Teil der Arbeit fokussiert eine kritische Analyse der methodischen Vorgehensweise der hier vorliegenden Hochschulstudien. Dabei erfolgt eine kritische Auseinandersetzung mit der Musterlösung als externes Außenkriterium sowie der empirischen Überprüfung des eingesetzten Instrumentes (T-MITOCAR) bei der Leistungsbewertung textbasierter Lernergebnisse.

Die Güte des in dieser Studie herangezogenen Instrumentes wurde in zahlreichen Studien getestet und kann als sehr gut eingeschätzt werden (vgl. Pirnay-Dummer, 2010; Pirnay-Dummer, Ifenthaler & Spector, 2010). Die Software ist jedoch für die richtige Verwendung von Grammatik- sowie Rechtschreiberegeln völlig blind. Dies sind jedoch Aspekte, die bei der Bewertung textbasierter Prüfungsleistungen von Relevanz sind. In der vorliegenden Studie fanden diese in der Bewertung keine Berücksichtigung, da ausschließlich nach inhaltlichen Kriterien bewertet wurde. Zur Überprüfung, wie ähnlich die einzelnen textbasierten Prüfungsleistungen dem Erwartungshorizont kommen (der textbasierten

Musterlösung bzw. dem Lehrtext), muss die valide Verwendung externer Referenzmodelle berücksichtigt werden. In der vorliegenden Untersuchung wurde dieses Kriterium auf Grundlage der Einschätzung eines Experten aus den gesuchten Gegenstandsbereichen berücksichtigt. Dennoch kann nicht ausgeschlossen werden, dass das Heranziehen mehrerer Experten möglicherweise zu einer valideren Musterlösung geführt hätte.

6.1.5.4 Auswertungsverfahren

Die statistische Auswertung der erhobenen Daten erfolgt mit der Software SPSS (IBM SPSS Statistics 20). Die deskriptiven Ergebnisse werden mittels des arithmetischen Mittels, Median, Standardabweichung und der Maximal- und Minimalwerte dargestellt. Vor jeder Datenauswertung findet eine Analyse der zugrunde gelegten Kriterien auf Normalverteilung mittels des *Kolmogorov-Smirnov-Tests* statt. Alle statistischen Auswertungen legen eine Irrtumswahrscheinlichkeit von $\rho = 0.05$ zugrunde.

6.2 Konzeptionen der Schuluntersuchungen

Der folgende Abschnitt stellt die Konzeptionen der Teilstudien im Schulkontext dar. Dies erfolgt der Übersicht halber nach Fächern getrennt. Die Überprüfung der technologiegestützten Leistungsbewertung durch T-MITOCAR im schulischen Kontext erfolgte am Beispiel von klassischen Leistungserfassungsmethoden (in Form von Klausuren und eines Aufsatzes). Diese Arbeiten umfassten ein naturwissenschaftliches Fach (Biologie), ein sprachliches Fach (Deutsch), ein gesellschaftliches Fach (Religion), sowie ein musisch-künstlerischen Fach (Kunst). Es sollte ein möglichst breites Spektrum in unterschiedlichen Unterrichtsfächern in Bezug auf die Textproduktion ermöglichen. Das Auswahlkriterium lag darin begründet, dass in diesen Fächern als reguläre Prüfungsleistung textproduzierende Aufgabenstellungen enthalten waren. Die Teilnahme an diesen Studien war sowohl von Lehrer- wie auch von Schülerseite freiwillig. Die Lehrkräfte erstellten die für sie ideale textbasierte Musterlösung um einen technologiegestützten Vergleich zu haben. Weiter sollte es einen Vergleich ermöglichen inwiefern sich diese durch eine Intervention verändern lässt. Diese Musterlösung wurde in einer Post-Hoc-Analyse auf Grundlage der vorliegenden Bewertungskriterien (die von den

Lehrkräften erstellt und zur Einschätzung der Schülerleistungen herangezogen wurden) dahin gehend eingeschätzt, inwiefern alle Kriterien darin abgedeckt waren.

Zunächst erfolgte ein quantitativer Zugang in dem die durch die Lehrkräfte vergebenen Punkte hinsichtlich der einzelnen Schülerergebnisse erfasst und mit den strukturellen und semantischen Ähnlichkeitswerten verglichen wurden. Dies sollte die Frage beantworten, inwiefern sich die Kriterien orientierte Bewertung der Lehrkräfte auf Grundlage deren selbst erstellten textbasierten Musterlösungen technologiegestützt abbilden lässt. Anschließend erfolgte ein qualitativer Zugang, indem die Bewertungskriterien der Lehrkräfte näher erfragt wurden sowie deren Bewertungsqualität mithilfe von Interviews erfasst wurden. Dies sollte zusätzliche Informationen über den Ausprägungsgrad der diagnostischen Expertise der an dieser Studie teilgenommenen Lehrkräfte liefern. Diese wurden anschließend auf Grundlage der Qualitativen Inhaltsanalyse ausgewertet (vgl. Mayring, 2008; Mayring & Gläser-Zikuda, 2008). Dies erfolgte mithilfe der induktiven sowie der deduktiven Analyse. Der folgende Abschnitt stellt die Umsetzung der empirischen Untersuchung in den einzelnen Unterrichtsfächern dar. Die Untersuchungen sind nach Fächern sortiert dargestellt (Biologie, Deutsch, Religion, Kunst).

6.2.1 Konkrete Bewertungssituation im Unterrichtsfach Biologie

Für diese Untersuchung wurde eine Biologieklausur herangezogen, die die Schüler im Laufe des Schuljahres schrieben. Sie bestand sowohl aus Aufgabenstellungen, in denen sie ihr Wissen in Form von Stichworten externalisierten, als auch aus einer textproduzierenden Aufgabenstellung. Für die Untersuchung wurde die Aufgabenstellung herangezogen, in der Schüler ihr Wissen textbasiert externalisierten. Die Aufgabenstellung war die folgende:

> *„Gibt es harmlose Drogen? Diese Fragestellung soll von Ihnen in einem Artikel für eine Schülerwandzeitung bearbeitet werden. Schließen Sie in Ihre Argumentation geeigneten Materialen aus den beigefügten Texten und mindestens ein weiteres Beispiel Ihrer Wahl ein. Der Umfang des Artikels für die Wandzeitung darf maximal zwei Seiten (Vorder- und Rückseite eines Blattes) betragen."*

Die beiden an dieser Untersuchung teilnehmenden Lehrkräfte im Unterrichtsfach Biologie wurden gebeten eine für sie ideale textbasierte Musterlösung zu erstellen. Jede Lehrkraft erstellte unabhängig von der anderen Lehrkraft die für sie zur Bewertung wichtig erscheinenden Kriterien und analysierte auf deren Grundlage die individuellen Schülerleistungen. Die Darstellung der von den Lehrkräften erstellten Kriterien folgt im Ergebnisbericht. Die beiden Lehrkräfte bewerteten dieselben Schülerergebnisse unabhängig voneinander. Dies sollte der Überprüfung der Frage dienen, inwiefern dieselben Schülerleistungen durch unterschiedliche Experten (in dem Fall die Lehrkräfte mit dem Expertisewissen aus dem gesuchten Gegenstandsbereich Biologie) unabhängig voneinander zu ähnlichen Bewertungen kommen.

6.2.2 Konkrete Bewertungssituation im Unterrichtsfach Deutsch

Im Laufe des Schuljahres schrieben die Schüler eine Erörterung, welche für die vorliegende Untersuchung herangezogen wurde. Hierzu erörterten sie die Fragestellung, ob soziale Netzwerke ein Ersatz für echte soziale Kontakte und reale Erfahrungen sind oder ob diese zur Selbstisolation führen. Die Schüler bezogen sich dabei auf einen Artikel aus der Zeitschrift

Focus sowie einzelne Cartoons und die Unterrichtsinhalte, die bereits im Vorfeld zu diesem Themengebiet erarbeitet wurden. Es gab folgende Aufgabenstellung:

> *„Sind soziale Netzwerke ein Ersatz für echte soziale Kontakte und reale Erfahrungen oder führen sie letztendlich zur sozialen Selbstisolation? Erörtere diese Frage unter Berücksichtigung des oben genannten Textausschnittes, der Cartoons und der Vorarbeiten aus dem Unterricht zu sozialen Netzwerken.“*

Die Lehrkraft bewertete die Schülertexte auf Grundlage der selbst erstellten Kriterien, die die Schüler bereits im Unterricht kennengelernt hatten, bevor die Klausur geschrieben wurde.

6.2.3 Konkrete Bewertungssituation im Unterrichtsfach Religion

Für dieses Fach wurde eine Klausur herangezogen, die im Laufe des Schuljahres geschrieben wurde. Sie bestand ausschließlich aus Fragen, die die Schüler in Form von Texten beantworteten. Es gab folgende Aufgaben:

1. *Fasse den Text von Franz Alt mit eigenen Worten zusammen!*
2. *Vergleiche, auch mithilfe des Textes von Franz Alt, Jesu Erwartung und die jüdische Erwartung des Reiches Gottes!*
3. *Deute das Gebot der Feindesliebe nach Matthäus 5,43-48 mithilfe einer im Unterricht behandelten Auslegungsart!*

Beide Lehrer entwickelten gemeinsam eine textbasierte Musterlösung sowie Kriterien für die Bewertung der von ihnen gemeinsam erstellten Aufgabenstellungen. Beide Lehrkräfte bewerteten die Schülertexte ihrer eigenen Klassen auf Grundlage der gemeinsam erstellten Kriterien.

6.2.4 Konkrete Bewertungssituation im Unterrichtsfach Kunst

Im Kunstunterricht wurde eine Klausur herangezogen, die im Laufe des Schuljahres von den Schülern geschrieben wurde. Die Aufgabenstellung ist die folgende:

Beschreibe, analysiere und interpretiere das in der Technik der Decalcomanie entstandene Gemälde „Die Versuchung des Heiligen Antonius“ (1945, Öl auf Leinwand, 109 x 129) anhand des Arbeitsblatts „Vorgehensweise in der Klausur“.
Achte beim Analyse-Teil darauf, auch etwas zur Technik der Decalcomanie bzw. der Frottage oder Grattage einzubringen, und beim Interpretations-Teil, etwas aus der Biografie von Max Ernst.

Die Lehrkraft erstellte eine textbasierte Musterlösung sowie Kriterien, auf deren Grundlage die einzelnen Schülerleistungen bewertet wurden.

6.2.5 Untersuchungsdesign

Die Untersuchungen im schulischen Kontext orientieren sich am Mixed-Method-Design (vgl. Teddlie & Tashakkorie, 2010). Dabei erfolgen quantitative sowie qualitative Zugänge. Die Lehrkräfte bewerten dieselben Schülerleistungen vor und nach der Intervention in Form von Noten bzw. Leistungspunkten (vgl. Abbildung 2). Um ein detaillierteres Verständnis über deren Bewertungen der Schülertexte zu erzielen, wurde ein qualitativer Zugang gewählt. Er erfolgte mithilfe eines leitfadengestützten Interviews vor und nach der Schulung. In Interviews wurden deren zugrunde gelegten Kriterien näher erfragt, was ein differenzierteres Bild über die zugrunde gelegten Kriterien ermöglichte. Schließlich wurde deren Vorgehensweise von der Klausurerstellung bis zur Notengebung erfragt um ein differenzierteres Bild über deren zugrunde gelegten Klausurfragen sowie deren Bewertung zu zeigen. Darüber hinaus wurde die Bewertungsqualität der Lehrkräfte erfragt. Dies erlaubte Rückschlüsse auf die Qualität der Leistungsbewertungen, die ohne diesen qualitativen Zugang nicht möglich gewesen wären.

Abbildung 2 Pre-Post-Untersuchungsdesign

	Quantitative Erhebung	**Qualitative Erhebung**
1. Messzeitpunkt (April/Mai 2011)	Bewertung der Schülertexte durch die Lehrkräfte	Interview
Intervention (Juni/ Juli 2011)		
2. Messzeitpunkt (Juni/Juli 2011)	Bewertung der Schülertexte durch die Lehrkräfte	Interview

6.2.6 Stichprobenbeschreibungen

Der folgende Abschnitt beschreibt die Stichproben der einzelnen Fächer.

6.2.6.1 Stichprobenbeschreibung im Unterrichtsfach Biologie

Die Stichprobe umfasst (N = 29) Schüler (3 männliche, 26 weibliche) aus insgesamt zwei Klassen. Das Durchschnittsalter beträgt 17.5 Jahre. Die erste Klasse umfasst (N = 22) Schüler, wovon alle weiblich sind und das Durchschnittsalter bei 17.5 liegt. Die zweite Klasse (N = 7) umfasst drei männliche und weibliche Gymnasialschüler. Das Durchschnittsalter liegt bei 17.4 Jahre.

6.2.6.2 Stichprobenbeschreibung im Unterrichtsfach Deutsch

Die Stichprobe umfasst (N = 17) (davon sechs männliche und elf weibliche) Gymnasialschüler der zehnten Klasse. Das Durchschnittsalter beträgt 16 Jahre.

6.2.6.3 Stichprobenbeschreibungen im Unterrichtsfach Religion

Im Unterrichtsfach Religion liegen zwei Teilstudien vor. Die Stichprobe der ersten Teilstudie umfasst (N = 15) (davon neun männliche und sechs weibliche) Gymnasialschüler der 12. Klasse. Das Durchschnittsalter beträgt 18.2 Jahre.
Die Stichprobe der zweiten Teilstudie umfasst N = 29 Schüler. Davon nahmen zwölf Schüler an dieser Teilstudie teil. Die Stichprobe besteht aus fünf männlichen und sieben weiblichen Schülern. Das Durchschnittsalter liegt bei 18.8 Jahren.

6.2.6.4 Stichprobenbeschreibung im Unterrichtsfach Kunst

Die Stichprobe umfasst (N = 32) (davon 11 männliche und 21 weibliche) aus Gymnasialschüler der 11 Klasse. Das Durchschnittsalter beträgt 17 Jahre. Die Stichprobe umfasst zwei Klassen, wovon die erste Klasse N = 15 (davon 6 weibliche und 9 männliche) Schüler umfasst und das Durchschnittsalter bei 17.7 Jahren liegt. Die zweite Klasse umfasst N = 17 (davon fünf weibliche und zwölf männliche) Schüler und das Durchschnittsalter beträgt 18.1 Jahre.

6.2.7 Methodisches Vorgehen

6.2.7.1 Datenerhebung

Es erfolgt sowohl ein quantitativer als auch ein qualitativer Zugang. Die durch die Lehrkräfte vergebenen Punkte werden quantitativ erfasst und deren herangezogenen Kriterien sowie deren Vorgehensweise und Bewertungsqualität werden qualitativ mithilfe von Interviews erfasst.

6.2.7.2 Quantitative Erhebung

Die Schüler bearbeiteten die Klausuraufgabenstellungen in den Klausuren in einer für sie gewöhnlichen Prüfungssituation. Dabei wurde sichergestellt, dass es für die Schüler eine gewöhnliche Situation darstellte, ihr Wissen in diesen Unterrichtsfächern textbasiert zu externalisieren. Die Bearbeitung der Aufgabenstellung erfolgte handschriftlich; dies sollte ungewohnte Nebeneffekte verhindern. Die Klausuren wurden digitalisiert und anonymisiert. Die Lehrkräfte bewerteten die individuellen Schülerleistungen in einer für sie gewohnten Situation, da sie auf ihre eigens erstellten Kriterien zurückgriffen. Ungewohnt für sie war die Erstellung der textbasierten Musterlösung. Für die vorliegenden Untersuchungen wurden die durch die Lehrkräfte vergebenen Leistungspunkte sowie deren ausformulierten Musterlösungen herangezogen.

6.2.7.3 Validitätsbestimmung der Musterlösungen

Die Musterlösungen wurden in einer Post-Hoc-Analyse daraufhin eingeschätzt, inwiefern alle, in den dazugehörigen Kriterien abgebildete Aspekte in den jeweiligen Musterlösungen enthalten sind. Dies sollte die Inhaltsvalidität der Musterlösungen hinsichtlich der Bewertungskriterien gewährleisten.

6.2.7.4 Qualitative Erhebung

Die Interviews fanden alle telefonbasiert statt. Dies hatte pragmatische Gründe und sollte zudem sicherstellen, dass mögliche Nebeneffekte ausgeschlossen werden, die entstehen könnten wenn die eine Lehrkraft „face-to-face“ und die andere telefonbasiert befragt werden.

6.2.7.5 Intervention

Nach der ersten Bewertungsphase erhielten die Lehrkräfte eine Schulung in Bezug auf ihre diagnostische Expertise. Diese umfasste Aspekte der Bewertungsqualität wie beispielsweise die Berücksichtigung der wissenschaftlichen Standards (Objektivität, Zuverlässigkeit sowie Gültigkeit) sowie das Minimieren von Fehlerquellen im Prozess der Bewertung. Dies sollte später gewährleisten, dass alle Lehrkräfte die individuellen Schülerleistungen auf Grundlage diagnostischer Expertise einschätzten. Zu Beginn der Schulung sowie unmittelbar am Ende des Trainings wurden die Lehrkräfte gebeten, aufzuschreiben, was aus Ihrer Perspektive eine faire Notengebung ausmacht. Die persönlichen Aufschriebe der Lehrkräfte fanden für die Auswertung keine Berücksichtigung, sondern dienten vielmehr dazu, mit den Lehrkräften zunächst ins Gespräch zu kommen. Im Laufe der Schulung erfuhren sie was sie tun können um die einzelnen (aus wissenschaftlicher Sicht) wichtigen Kriterien für eine faire Bewertung zu berücksichtigen. Dies betraf beispielsweise die Erstellung der Klausur, bevor deren Inhalte im Unterricht behandelt wurden (hinsichtlich der Inhaltsvalididät), sowie die Vermeidung von Fehlerquellen im Prozess der Bewertung, indem beispielsweise die Namen abgedeckt wurden.

6.2.7.6 Erhebungsinstrumente

Die Entwicklung des Leitfadeninterviews (siehe Anhang A) erfolgte in Anlehnung an Helmke (2009) theoriegeleitet. Zudem wurden wie bereits im Hochschulkontext die Instrumente T-MITOCAR und AKOVIA zur Bestimmung der semantischen sowie strukturellen Ähnlichkeiten zwischen den textbasierten Musterlösungen der Lehrkräfte sowie den einzelnen Schülertexten herangezogen.

6.2.7.7 Methodenkritische Anmerkungen

Die Güte der technologiebasierten Instrumente wurde bereits dargestellt. Die Güte des Kodierleitfadens, welcher zur deduktiven Analyse herangezogen ist in Abschnitt 8.6.4 und 8.6.5 dargestellt.

6.2.7.8 Auswertungsverfahren

Die Auswertung der quantitativen Daten erfolgt mit der Software SPSS (IBM SPSS Statistics 20) (siehe Kapitel Auswertungsverfahren Hochschule).

6.2.7.9 Leitfadeninterview und Interviewauswertung

Die Erfassung der Bewertungsqualität der Lehrkräfte erfolgte mithilfe eines leitfadengestützten Interviews. In diesem wurde in Anlehnung an Helmke (2009) deren diagnostische Expertise erfasst. Zudem wurden die durch die Lehrkräfte bestimmten Kriterien sowie deren konkrete Umsetzung erfragt sowie deren Vorgehensweise vom Erstellen der Klausur bis hin zur Benotung. Die Auswertung der Interviews erfolgte mithilfe der Qualitativen Inhaltsanalyse (vgl. Mayring, 2008 b; Mayring & Gläser-Zikuda, 2008). Dabei erfolgten eine theoriegeleitete (deduktive) Kategorienbildung und eine Skalierung (auf Ordinalebene) der Lehreraussagen bezüglich ihrer Bewertungsqualität. Dies ermöglichte es später, die Lehrkräfte hinsichtlich ihrer diagnostischen Fähigkeiten zu klassifizieren und Rängen zuzuordnen. Im nächsten Schritt erfolgte eine induktive Auswertung der Kriterien sowie der Vorgehensweisen der Lehrkräfte.

6.2.8 Leitfadeninterview

Die telefonbasierten Interviews mit den Lehrkräften fanden jeweils unmittelbar nach deren Bewertung der textbasierten Schülerleistungen statt. Diese orientierten sich an Leitfragen in Bezug auf deren Bewertungskriterien, Vorgehensweise sowie deren Bewertungsqualität (Beachtung der Gütekriterien, Minimierung der Fehlerquellen, Berücksichtigung unterschiedlicher Aufgabenniveaus sowie Bezugsnormen). Folgender Auszug aus dem Interviewleitfaden verdeutlicht dies (siehe Anhang A).

Bewertungskriterien:

- Nach welchen Kriterien bewerten Sie textbasierte Schülerleistungen?

Berücksichtigung der Gütekriterien:

- Welche Hinweise geben Sie Ihren Schülern beim Austeilen der Klausuren? Wie reagieren Sie auf Nachfragen?
  (Durchführungsobjektivität)

- Welche Bewertungskriterien kennen Ihre Schüler, bevor die Klausur geschrieben wird? *(Auswertungsobjektivität)*
- Haben Sie schon mal versucht, mehrere Bewerter heranzuziehen? Beispielsweise in „schwierigen" Fällen. Wie sind Sie dabei vorgegangen? *(Reliabilität*

Minimierung der Bewertungsfehler:

- Wie minimieren Sie Bewertungsfehler?

Bezugsnorm:

- An welchem Maßstab orientieren Sie sich bei der Bewertung der Schülerleistungen?

Im Zusammenhang mit den Bewertungskriterien, erfolgte beim zweiten Messzeitpunkt zudem die Frage, wie genau man bei wiederholter Bewertung vorgehen müsste, um die in den Kriterien abgebildeten Vorgaben in den einzelnen Schülertexten wiederzufinden

6.2.9 Auswertung der Interviews mittels Qualitativer Inhaltsanalyse

Die telefonbasierten Interviews wurden transkribiert und anschließend mit induktiver und deduktiver Inhaltsanalyse analysiert. Dabei wurden die ersten beiden Fragestellungen mit induktiver Analyse beantwortet. Diese umfassten die Bewertungskriterien, welche die Lehrkräfte zugrunde legten sowie deren Vorgehensweise von der Klausurerstellung bis hin zur Bewertung. Im nächsten Schritt wurde die Fragestellung in Hinsicht auf deren Bewertungsqualität deduktiv untersucht. Dabei analysierten zwei Gutachter (unabhängig voneinander) die Interviews in Anlehnung an das zuvor festgelegte Regelwerk. Vor der Analyse der Interviews wurden die Gutachter in Bezug auf die Methode der Qualitativen Inhaltsanalyse sowie des Konstrukts der Bewertungsqualität geschult. Dies sollte mögliche Messfehler im Umgang mit der Methode sowie der Erfassung des Konstrukts (Bewertungsqualität) vermeiden. Ziel der induktiven Analysen war es, einen tieferen Einblick in die zugrunde gelegten Kriterien zu bekommen sowie ein elaborierteres Verständnis über die Vorgehensweise der Lehrkräfte beim Erstellen und Bewerten von Schülertexten. Ziel der deduktiven Interviewanalysen war es zunächst die zuvor festgelegten Kategorien bezüglich deren Bewertungsqualität (diagnostische Expertise) zu identifizieren um schließlich deren Ausprägungsgrad auf einer dreistufigen (ordinalskalierten) Skala zu bestimmen. Dabei orientiert sich

diese Analyse an der skalierenden Strukturierung. Für diese Einschätzungen wurde im Vorfeld ein Kodierleitfaden entwickelt (vgl. Tabelle 5, sowie Anhang B). Die (N = 2) Bewerter analysierten die Lehreraussagen auf Grundlage des Kodierleitfadens auf einer dreistufigen Skala. (0 = niedrige Ausprägung, 1 = mittlerer Ausprägung, 2 = starke Ausprägung).

6.2.9.1.1 Interkoderreliabilität

Die Zuverlässigkeit der deduktiven Inhaltsanalyse wurde in der vorliegenden Studie mithilfe der Interkoderreliabilität sowie dem Reliabilitätskoeffizienten nach Krippendorf berechnet (vgl. Mayring, 2008 b, S.113; Bortz & Döring, 2003).

6.2.9.1.2 Induktive Auswertung

Die Analyse der (N = 12) Interviews erfolgte zunächst unabhängig durch jeweils zwei Kodiere. Diese analysierten die Textstellen und fassten diese zunächst in (nicht vorgegebene) Kategorien. Anschließend fassten sie die induktiv gebildeten Kategorien zu (theoriegeleiteten) Oberkategorien zusammen. Für die Bestimmung der Messzuverlässigkeit wurden schließlich nur die gemeinsamen Oberkategorien als gleich kategorisiert, deren Kategorien bereits als gleich kategorisiert worden waren. Wenn sich in der Beschreibung der Kategorie von einem Kodierer Inhalte fanden, die in der Kategorie des anderen Kodierers als Kategorie bestimmt wurden, so wurde dies nicht als dieselbe Kategorie definiert. Danach fassten die jeweiligen beiden Bewerter in einem gemeinsamen Gespräch ihre (unabhängig voneinander) erstellen Kategorien und Oberkategorien zu einem gemeinsamen Konsens an Kategorien zusammen.

6.2.9.1.3 Deduktive Auswertung

Die Analyse der (N = 12) Interviews erfolgte unabhängig durch jeweils zwei Kodierer. Diese analysierten die Textstellen zunächst mithilfe des Kodierleitfadens und schätzten diese anschließend nach ihrem Ausprägungsgrad ein. Für die Auswertung wurden schließlich nur die Kategorien herangezogen, die von beiden Bewertern gleich kategorisiert wurden. Dabei wurde von beiden markierten Textstellen, jeweils die gemeinsame Schnittmenge herangezogen. Im nächsten

Schritt wurde die Interkoderreliabilität für die eingeschätzten Ausprägungsgrade berechnet.

Tabelle 5 Ausschnitt aus dem Kodierleitfaden

Variable	Ausprägung	Definition	Ankerbeispiel	Kodierregeln
Durchführungs-objektivität	Starke Ausprägung	Die Lehrkraft achtet darauf, dass die Schüler unter gleichen Rahmenbedingungen schreiben. Sie nennt Maßnahmen und konkrete Beispiele, um dies zu gewährleisten.	„Ich sage meinen Schülern vor der Klausur, welche Hilfsmittel sie verwenden dürfen. Dies tue ich gewöhnlich schon bei der Bekanntgabe des Klausurtermins und nicht erst zur Klausur.“	Mindestens 1 Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „niedrige Ausprägung“ enthalten sein.
	Mittlere Ausprägung	Die Lehrkraft achtet darauf, dass die Schüler unter gleichen Rahmenbedingungen schreiben. Sie nennt keine Maßnahmen und keine konkreten Beispiele, um dies zu gewährleisten.	„Ja, die schreiben unter den gleichen Bedingungen.“	Sobald über diesen Aspekt gesprochen wird, ist er nicht mehr stark oder niedrig ausgeprägt.
	Niedrige Ausprägung	Die Lehrkraft achtet nicht darauf, dass die Schüler unter gleichen Rahmenbedingungen schreiben.	„Darauf achte ich nicht.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „starke Ausprägung“ enthalten sein.

7 Ergebnisse der Hochschuluntersuchungen

Die Darstellung der empirischen Ergebnisse der beiden Untersuchungen im Hochschulkontext folgt in chronologischer Reihenfolge. Diese schließen zunächst eine deskriptive und daran anschließende eine hypothesenprüfende Darstellung der Befunde ein.

7.1 Ergebnisse der ersten Hochschuluntersuchung

Aus den Gegenstandsbereichen des selbstregulierten Lernens und der Metakognition verfassten die Probanden Texte. Bei der Beantwortung der Frage zum selbstregulierten Lernen konnten von insgesamt N = 45 Klausuren nur N = 40 analysiert werden, weil in den anderen Fällen entweder keine Texte produziert wurden (in N = 3 Fällen), oder die Texte für eine modellbasierte Analyse mit T-MITOCAR zu klein waren (in N = 2 Fällen). Bei der Aufgabenstellung zur Metakognition konnten von insgesamt N = 45 Klausuren nur N = 37 analysiert werden, weil in den anderen Fällen keine Texte verfasst wurden. Ein möglicher Grund hierfür könnte sein, dass die Lernenden sich in dem bestimmten Themengebiet nicht vorbereitet hatten.

7.1.1 Ergebnisse

Die Deskription der Ergebnisse orientiert sich an den im fünften Kapitel formulierten Fragestellungen. Es erfolgt eine Darstellung der in Form von Punkten gegebenen Leistungseinschätzungen der beiden Dozenten (vgl. Tabelle 6). Weiter eine Darstellung der Ähnlichkeitskennwerte der Lernergebnisse mit den jeweiligen Außenkriterien, die mithilfe von T-MITOCAR (vgl. Tabelle 7 und 8) zustande kamen. Die deskriptiven Ergebnisse in Tabelle 6 verdeutlichen eine ähnliche Einschätzung der beiden Bewerter. Insgesamt wurde das von den Studierenden in Form von Texten externalisierte Wissen bei der Aufgabenstellung des selbstregulierten Lernens von beiden Bewertern als höher eingeschätzt als die produzierten Texte bezüglich der Metakognition.

Tabelle 6 Deskription der Bewertungen

	Selbstreguliertes Lernen (N = 42)				Metakognition (N = 42)			
	AM	SD	MD	Range	AM	SD	MD	Range
1. Bewerter	6.87	2.93	7.25	0-11.00	3.88	2.48	4.00	0-10.50
2. Bewerter	6.90	2.89	7.50	0-12.00	3.74	2.16	3.50	0-8.50

Die folgenden Tabellen (vgl. Tabelle 7 und 8) zeigen die Ähnlichkeitswerte, welche die Nähe der Lernergebnisse zu den jeweiligen Referenzmodellen abbilden. Als Referenzmodelle wurden sowohl die Musterlösung herangezogen, als auch das Gesamtmodell, welches alle Lernergebnisse der Studierenden umfasst, sowie der Lehrtext. Da in einzelnen Antworten (N = 5 beim selbstregulierten Lernen und N = 2 bei der Metakognition) ausschließlich Stichpunkte enthalten waren, wurden diese aus dem Gesamtmodell entfernt. Die Texte ähnelten den jeweiligen Referenztexten strukturell mehr als inhaltlich. Dies betrifft beide Aufgabenstellungen wobei die Texte zum selbstregulierten Lernen den Vergleichstexten strukturell noch mehr ähnelten als die zur Metakognition. Das Graphical Matching wies dabei eine besonders hohe Übereinstimmung auf (siehe Tabelle 7). Dies lässt vermuten, dass die Studierenden hier ein ähnlich breites und komplexes Wissen hatten wie das in den jeweiligen Vergleichstexten.

Tabelle 7 Deskription der Ähnlichkeitskennwerte zum selbstregulierten Lernen (N = 40)

		ML		GMM		GMO		LT	
	Kennwerte	AM	SD	AM	SD	AM	SD	AM	SD
Struktur	Surface Matching	0.50	0.26	0.43	0.25	0.44	0.25	0.49	0.25
	Graphical Matching	0.69	0.26	0.69	0.26	0.64	0.25	0.68	0.22
	Structural Matching	0.24	0.23	0.36	0.36	0.36	0.36	0.41	0.35
	Gamma Matching	0.60	0.25	0.50	0.23	0.50	0.23	0.60	0.25
Semantik	Concept Matching	0.14	0.09	0.28	0.12	0.28	0.13	0.11	0.08
	Propositional Matching	0.02	0.04	0.07	0.08	0.07	0.08	0.00	0.01
	Balanced Semantic Matching	0.08	0.17	0.22	0.20	0.21	0.19	0.02	0.08

ML = Musterlösung; GMM = Gesamtmodell mit Stichpunkten; GMO = Gesamtmodell ohne Stichpunkte; LT = Lehrtext

Die verfassten Texte zur Metakognition zeigten insbesondere bei Gamma hohe Übereinstimmungen mit den jeweiligen Referenzmodellen, was darauf hinweist, dass die Modelle der Studierenden im Vergleich zu den jeweiligen Referenzmodellen eine ähnliche Knotendichte aufweisen (vgl. Tabelle 8).

Tabelle 8 Deskription der Ähnlichkeitskennwerte zur Metakognition (N = 37)

		ML		GMM		GMO		LT	
	Kennwerte	AM	SD	AM	SD	AM	SD	AM	SD
Struktur	Surface Matching	0.33	0.26	0.30	0.30	0.31	0.30	0.30	0.30
	Graphical Matching	0.51	0.26	0.51	0.26	0.51	0.26	0.49	0.25
	Structural Matching	0.26	0.34	0.22	0.31	0.21	0.30	0.20	0.30
	Gamma Matching	0.58	0.22	0.52	0.26	0.55	0.25	0.55	0.25
Semantik	Concept Matching	0.12	0.10	0.24	0.12	0.23	0.12	0.23	0.12
	Propositional Matching	0.01	0.03	0.06	0.06	0.08	0.06	0.07	0.06
	Balanced Semantic Matching	0.06	0.14	0.20	0.19	0.23	0.21	0.23	0.21

ML = Musterlösung; GMM = Gesamtmodell mit Stichpunkten; GMO = Gesamtmodell ohne Stichpunkte; LT = Lehrtext

7.1.2 Qualität der Bewertungskriterien

Die Qualität der Bewertungskriterien wurde im ersten Schritt mithilfe von Cohen's Kappa überprüft. Danach folgte eine Untersuchung der einzelnen Skalen mithilfe des Cronbach's Alpha. Und zum Schluss wurde mithilfe der rangskalierten Interkoderreliabilität geprüft, inwiefern die beiden Hochschuldozenten dieselben Kriterien in den Texten fanden.

7.1.2.1 Selbstreguliertes Lernen

Die einzelnen Kriterien des selbstregulierten Lernens können bis auf die Definition in der ersten Phase, die Motivation und die Lernziele sowie die Probleme und Grenzen, als gut eingeschätzt werden (vgl. Bühner, 2004, S. 129). Das Kriterium „Post-Aktionale-Phase" ergab dabei eine hohe Übereinstimmung (vgl. Tabelle 9).

Tabelle 9 Bewertungsübereinstimmung der Kriterien zum selbstregulierten Lernen

Kriterium (Allgemein)	Kriterium (spezifisch)	Cohen's Kappa
Definition	Definition	0.53 (*)
Präaktionale Phase	Präaktionale Phase	0.90 (*)
	Emotionen	0.63 (*)
	Motivation	0.53 (*)
	Lernziele	0.51 (*)
	Planung	0.70 (*)
Aktionale Phase	Aktionale Phase	0.90 (*)
	Lernstrategien	0.67 (*)
	Self-Monitoring	0.86 (*)
Post-Aktionale Phase	Post-Aktionale Phase	0.95 (*)
	Selbst-Reflexion	0.64 (*)
	Konsequenzen	0.63 (*)
Probleme	Probleme und Grenzen	0.42 (*)

Alle durch ein (*) gekennzeichneten Ergebnisse sind signifikant.

Die Überprüfung der Messgenauigkeit (*Reliabilität*) der einzelnen Skalen zeigte auf Grundlage des Cronbach Alpha-Wertes auf allen Skalen eine niedrige Genauigkeit (ebd., S. 129). Dies traf auf die Bewertungen beider Hochschuldozenten zu (vgl. Tabelle 10). Die Daten sind auf Grundlage des Kolmogorov-Smirnov-Tests normalverteilt.

Tabelle 10 Skalenüberprüfung des selbstregulierten Lernens mithilfe des Cronbach's Alpha-Wertes

Kriterium (Allgemein) –Skalen	α (1 Bewerter)	α (2 Bewerter)	Kolmogorov-Smirnov-Test
Präaktionale Phase	0.65	0.59	normalverteilt
Aktionale Phase	0.47	0.42	normalverteilt
Post-Aktionale Phase	0.68	0.63	normalverteilt
Gesamtskala	0.78	0.79	normalverteilt
Gesamtskala (ohne Definition und Grenzen)	0.85	0.80	normalverteilt

Die Untersuchung der Interkoderreliabilitäten der einzelnen Skalen ergab eine hohe Bewertungsübereinstimmung bei den einzelnen Phasen (Präaktionale Phase, Aktionale Phase sowie Post-Aktionale Phase). Eine niedrige Messgenauigkeit zeigte sich bei der Definition und eine gute Übereinstimmung zeigte sich bei den Problemen und Grenzen des Modells (vgl. Tabelle 11). Die Untersuchung der Gesamtskala ergab ebenso eine hohe Übereinstimmung (ebd., S. 129).

Tabelle 11 Bewertungsübereinstimmung der einzelnen Skalen

Kriterium (Allgemein) – Skalen	Interkoderreliabilität
Definition (Singleitem)	0.73 (*)
Präaktionale Phase	0.91 (*)
Aktionale Phase	0.94 (*)
Post-Aktionale Phase	1.00 (*)
Probleme und Grenzen (Singleitem)	0.82 (*)
Gesamtskala	0.93 (*)

7.1.2.2 Metakognition

Die Untersuchung der Auswertungsobjektivität mithilfe des Cohen's Kappas ergab bei allen Kriterien eine niedrige Übereinstimmung zwischen den beiden Hochschuldozenten (vgl. Tabelle 12). Ein möglicher Grund hierfür könnte die unpräzise Beschreibung innerhalb der Kriterien gewesen sein.

Tabelle 12 Bewertungsübereinstimmung der Kriterien zur Metakognition

Kriterium	Cohen's Kappa
Definition	0.39 (*)
Metakognitives Wissen	0.23 (*)
Prozedurales Wissen	0.37 (*)
Selbst-Wirksamkeit	0.57 (*)
Selbst-Management	0.46 (*)
Selbst-Bewertung	0.38 (*)
Lernstrategien	0.22 (*)
Metakognitive Strategien	0.08
Strategische Lerner	0.22 (*)
Strategische Wissensnutzer	0.38 (*)
Selbstreguliertes Lernen	0.22
Metakognitives Lernen	0.10
Metakognitives Wissen	0.14

Die Überprüfung der Messgenauigkeit (Reliabilität) der Gesamtskala ergab mithilfe des Cronbach's Alpha- Wertes eine gute Messgenauigkeit bei den Bewertungen des ersten Hochschuldozenten und eine niedrige Messgenauigkeit bei den Bewertungen des zweiten Hochschuldozenten (vgl. Tabelle 13). Die Daten sind auf Grundlage des Kolmogorov-Smirnov-Tests normalverteilt.

Tabelle 13 Skalenüberprüfung zur Metakognition mit Hilfe des Cronbach's Alpha-Wertes

Kriterium	α 1 Bewerter	α 2 Bewerter	Kolmogorov-Smirnov-Test
Metakognition (Gesamt)	0.87	0.79	Normalverteilt

Die Untersuchung der Interkoderreliabilität (vgl. Tabelle 13) ergab eine niedrige Bewertungsübereinstimmung zwischen den beiden Begutachtern bei der Gesamtskala (Bühner, 2004, S. 129).

Tabelle 14 Bewertungsübereinstimmung der Gesamtskala

Kriterien Insgesamt	Interkoderreliabilität
Metakognition	0.73 (*)

7.1.3 Hypothesenprüfende Darstellung

Der folgende Abschnitt beinhaltet die Überprüfung der im Kapitel fünf aufgestellten statistischen Hypothesen. Die Überprüfung, inwiefern die beiden Hochschuldozenten auf Grundlage derselben Kriterienkataloge (vgl. Tabellen 1 - 2) zu ähnlichen Gesamteinschätzungen gelangen, sind bereits im vorangegangen Abschnitt 7.1.2 dargestellt. Dabei zeigte sich bei der ersten Aufgabenstellung eine hohe und bei der zweiten Aufgabenstellung eine gute Bewertungsübereinstimmung.

Zur Überprüfung der Hypothesen, ob sich die (inhaltlich orientierten) Bewertungen der Lehrkräfte durch die (semantischen) T-MITOCAR Kennwerte abbilden lassen, erfolgte zunächst eine nach strukturellen und anschließend nach semantischen Kennwerten getrennte Regressionsanalyse. Daran anschließend wurde eine Überprüfung der Korrelationen nach Spearman durchgeführt. Als Außenkriterien werden sowohl die Musterlösung als auch das Lernergesamtmodell und der Lehrtext herangezogen. Um die Darstellung überschaubar zu halten, werden nur die signifikanten Regressionsanalysen tabellarisch dargestellt. Die nicht-signifikanten Ergebnisse können dem Anhang entnommen werden (vgl. Anhang C). Ebenso der Übersichtlichkeit halber werden die Regressionen nicht für jeden Hochschuldozenten einzeln dargestellt, sondern in Bezug auf die Mittelung der Einschätzungen beider Dozenten. Dies kann in der bereits dargestellten ähnlichen Gesamteinschätzung begründet werden.

7.1.3.1 Selbstreguliertes Lernen

Die Regressionsanalysen (vgl. Tabelle 15) zeigen, dass die strukturellen Kennwerte nicht signifikant mit den inhaltlich orientierten Bewertungen beider Hochschuldozenten korrelieren. Dies entspricht den theoretischen Vorannahmen. Die inhaltlich orientierten Bewertungen beider Hochschuldozenten korrelieren signifikant mit den semantischen Kennwerten in Bezug auf das Lernermodells. Das Lernermodell umfasst alle zugrunde gelegten Lernergebnisse und spiegelt somit auch die soziale Bezugsnorm wieder. Dies trifft sowohl auf das Lernermodell zu, welches auch stichpunktartige Lösungen enthielt, als auch auf das Lernermodell in dem keine stichpunktartigen Lösungen enthalten waren. Demzufolge ließen sich die Einschätzungen der Leistungspunkte in Hinsicht auf die soziale Bezugsnorm auf Grundlage der Ähnlichkeitskennwerte vorhersagen. Die zugrunde gelegte Musterlösung sowie der Lehrtext korrelierten bei beiden Hochschuldozenten nicht signifikant mit den semantischen Kennwerten (siehe Anhang C, Tabelle 100 - 101). Deswegen dürfen mögliche signifikante Korrelationskoeffizienten hier nicht interpretiert werden. Demzufolge lassen sich auch die Leistungspunkte bei der kriterialen Bezugsnorm - (Lehrtext sowie Musterlösung) -nicht vorhersagen. Die erste aufgestellte Hypothese (H_1) darf in Bezug auf das Gesamtmodell angenommen werden – nicht jedoch in Bezug auf die Musterlösung und den Lehrtext.

Tabelle 15 Regressionsanalyse beider Bewerter

			GMM			**GMO**		
Kriterien	Kennwerte	df	Korr. R^2	ρ	F	Korr. R^2	ρ	F
Inhalt	Struktur	4	-0.03	0.59	0.71	0.04	0.25	1.40
Inhalt	Semantik	3	0.34	0.00	7.64	0.39	0.00	9.14

Die Überprüfung der Korrelationskoeffizienten (nach Spearman) (vgl. Tabelle 16) ergab bei der sozialen Bezugsnorm (Gesamtmodell) bei beiden Hochschuldozenten einen signifikanten Zusammenhang des Kennwertes Concept Matching sowie des Propositional Matching. Dies verdeutlicht, dass wenn die Studierenden die Begrifflichkeiten in einem ähnlichen Kontext wie dem der Musterlösung verwendeten, sie mehr Punkte in der Gesamtbewertung erzielten.

Tabelle 16 Korrelationskoeffizienten beider Bewerter

	Kennwerte	GM	GMO
Struktur	Surface Matching	0.17	0.18
	Graphical Matching	- 0.09	- 0.13
	Structural Matching	0.08	0.08
	Gamma Matching	- 0.01	- 0.01
Semantik	Concept Matching	0.58 (*)	0.61 (*)
	Propositional Matching	0.37 (*)	0.34 (*)
	Balanced Semantic Matching	0.20	0.17

7.1.3.2 Metakognition

Die Regressionsanalysen zeigen, dass keine strukturellen Kennwerte signifikant mit den inhaltlich orientierten Bewertungen beider Hochschuldozenten korrelieren. Dies entspricht den theoretischen Vorannahmen. Die inhaltlich orientierten Bewertungen korrelierten ebenfalls (entgegen der theoretischen Vorannahmen) nicht signifikant mit den semantischen Kennwerten (vgl. Tabelle 104). Demzufolge lassen sich die Leistungspunkte nicht durch die Ähnlichkeitskennwerte vorhersagen. Die erste Nullhypothese (H_{01}) muss beibehalten werden. Es besteht kein Zusammenhang (mithilfe der Regressionsanalysen und des korrigierten R^2) zwischen den inhaltlich orientierten Bewertungen (die durch die Hochschuldozenten ermittelt wurden) und den durch T-MITOCAR ermittelten semantischen Kennwerten.

7.1.4 Post-Hoc-Analyse

In einer Post-Hoc-Analyse wurden die Korrelationen zwischen den vergebenen Leistungspunkten und der Wortanzahl untersucht. Dies sollte versichern, dass die Hochschuldozenten nur die Lernergebnisse und keine Textlänge bewerteten; längere Texte sollten also nicht bevorteilt werden. Die deskriptive Betrachtung der durchschnittlichen Wortanzahl zeigte, dass die Texte hinsichtlich der Metakognition wesentlich kürzer ausfielen als die des selbstregulierten Lernens (vgl. Tabelle 17).

Tabelle 17 Deskription der Wortanzahl

	Selbstreguliertes Lernen (N = 45)				**Metakognition (N = 45)**			
	AM	SD	MD	Range	AM	SD	MD	Range
Wortanzahl	248.16	140.09	255	0-454	113.40	72.97	121	0-268

Die Analyse zeigte einen signifikanten Korrelationskoeffizienten (Spearman) bei der Textlänge sowie des Gesamteindrucks. Dies verdeutlicht Tabelle 18.

Tabelle 18 Wortanzahl und Leistungspunkte

	Selbstreguliertes Lernen	**Metakognition**
Bewerter	R	r
Dozent 1	0.64 (*)	0.47 (*)
Dozent 2	0.62 (*)	0.63 (*)

7.2 Ergebnisse der zweiten Hochschuluntersuchung

Die Gegenstandsbereiche aus denen die Probanden ihr Wissen in Textform externalisieren, umfassten wahlweise entweder das selbstregulierte Lernen oder die Lernstrategien. Dabei wählten N = 70 Studierende die Aufgabenstellung zu den Lernstrategien und N = 103 Studierende die Aufgabenstellung zum selbstregulierten Lernen. N = 10 Studierenden externalisierten weder zur Aufgabenstellung des selbstregulierten Lernens noch der Lernstrategien ihr Wissen. Ein möglicher Grund hierfür könnte gewesen sein, dass sie sich auf beide Themenbereiche nicht vorbereitet hatten. Von den insgesamt N = 103 Klausuren (bei der Aufgabenstellung zum selbstregulierten Lernen) konnten nur N = 99 mit den Referenzmodellen verglichen werden, weil in den anderen Fällen keine Texte vorlagen. Von den insgesamt N = 70 Klausuren (bei der Aufgabenstellung zu den Lernstrategien) konnten nur N = 57 mit den Referenzmodellen verglichen werden, weil in den anderen Fällen keine Texte vorlagen.

7.2.1 Ergebnisse

Die deskriptiven Ergebnisse in Tabelle 19 zeigen eine ähnliche Bewertung der beiden Hochschuldozenten. Insgesamt bewerteten diese die Texte des selbstregulierten Lernens als qualitativ besser als die Texte der Lernstrategien. Bei der ersten Aufgabenstellung gab es insgesamt 12.25 Punkte und bei der zweiten insgesamt 15 Punkte zu erreichen.

Tabelle 19 Deskription der Bewertungen

	Selbstreguliertes Lernen (N = 103)				**Lernstrategien (N = 70)**			
	AM	SD	MD	Range	AM	SD	MD	Range
1. Bewerter	5.80	2.19	6.25	0 - 9.63	4.96	3.21	4.25	0 - 14.50
2. Bewerter	5.88	2.36	6.38	0 - 9.88	4.09	2.73	3.80	0 - 12.60

MD = Median

Die folgende Tabelle (vgl. Tabelle 20) stellt die Ähnlichkeitswerte dar, welche die Nähe der Lernergebnisse mit den Referenzmodellen abbildet. Die strukturellen Ähnlichkeiten der Lernergebnisse mit den Musterlösungen sind größer als die semantischen Ähnlichkeitskennwerte. Dies trifft auf beide Aufgabenstellungen zu.

Bei der Aufgabenstellung des selbstregulierten Lernens weisen insbesondere die Kennwerte Graphical Matching sowie Gamma hohe Ähnlichkeiten auf, was zu der Annahme führt, dass die Studierenden bei beiden Gegenstandsbereichen ein ähnlich umfangreiches und komplexes Wissen hatten sowie eine ähnliche Knotendichte im Vergleich zu den jeweiligen Referenzmodellen.

Tabelle 20 Deskription der Ähnlichkeitskennwerte zum selbstregulierten Lernen (N = 99)

	Kennwerte	**ML**		**GMM**		**GMO**		**LT**	
		AM	SD	AM	SD	AM	SD	AM	SD
Struktur	Surface Matching	0.50	0.29	0.39	0.26	0.40	0.26	0.35	0.28
	Graphical Matching	0.63	0.24	0.56	0.25	0.56	0.25	0.48	0.25
	Structural Matching	0.26	0.26	0.33	0.35	0.35	0.36	0.35	0.36
	Gamma Matching	0.61	0.25	0.48	027	0.51	0.26	0.59	0.27
Semantik	Concept Matching	0.15	0.08	0.28	0.13	0.29	0.13	0.11	0.09
	Propositional Matching	0.02	0.05	0.09	0.09	0.09	0.09	0.01	0.02
	Balanced Semantic Matching	0.09	0.19	0.26	0.21	0.26	0.21	0.04	0.09

ML = Musterlösung; GMM = Gesamtmodell mit Stichwörtern; GMO = Gesamtmodell ohne Stichwörter; LT = Lehrtext

Die Betrachtung der Ähnlichkeitskennwerte (vgl. Tabelle 21) bei den Lernstrategien veranschaulichen, dass auf struktureller Ebene insbesondere das Graphical Matching hohe Übereinstimmungen aufweist. Dies deutet darauf hin, dass die Studierenden ein ähnlich breites und komplexes konzeptuelles Wissen hatten wie das in den jeweiligen Referenzmodellen.

Tabelle 21 Deskription der Ähnlichkeitskennwerte zu den Lernstrategien (N = 57)

	Kennwerte	**ML**		**GMO**		**LT**	
		AM	SD	AM	SD	AM	SD
Struktur	Surface Matching	0.36	0.25	0.37	0.26	0.37	0.26
	Graphical Matching	0.60	0.23	0.59	0.19	0.59	0.19
	Structural Matching	0.26	0.25	0.27	0.35	0.29	0.36
	Gamma Matching	0.51	0.33	0.50	0.33	0.37	0.27
Semantik	Concept Matching	0.13	0.12	0.20	0.09	0.09	0.08
	Propositional Matching	0.03	0.08	0.06	0.07	0.01	0.02
	Balanced Semantic Matching	0.11	0.22	0.21	0.21	0.04	0.10

ML = Musterlösung; GMO = Gesamtmodell ohne Stichwörter; LT = Lehrtext

Die Analysen des Gesamtmodells, in dem die Klausuren mit enthalten waren, in denen die Studierenden stichpunktartig antworteten, waren mit T-MITOCAR nicht möglich. Deswegen wurden für die Analysen nur die Klausuren in das Gesamtmodell aufgenommen, welche ausformulierte Sätze beinhalteten.

7.2.2 Qualität der Bewertungskriterien

Der folgende Teil der Arbeit beinhaltet die Analyse der Qualität der Bewertungskriterien. Hierfür erfolgt zunächst eine Untersuchung der Cohen's Kappa Werte. Daran anschließend erfolgt die Darstellung der Qualität der einzelnen Skalen mithilfe des Cronbach's Alpha. Und zum Schluss erfolgt die Überprüfung, inwiefern beide Hochschuldozenten auf Grundlage derselben Kriterien zu ähnlichen Bewertungen kamen.

7.2.2.1 Selbstreguliertes Lernen

Bei den einzelnen Kriterien des selbstregulierten Lernens ergab sich insgesamt eine niedrige Übereinstimmung zwischen den beiden Hochschuldozenten (vgl. Tabelle 22). Die Cohen's Kappa-Werte für die Kriterien: präaktionale Phase, Aktionale Phase, Post-Aktionale-Phase sowie die Emotionen und Motivation innerhalb der ersten Phasen können jedoch als gut eingeschätzt werden (vgl. Bühner, 2004, S. 129).

Tabelle 22 Bewertungsübereinstimmung der Kriterien zum selbstregulierten Lernen

Kriterium (Allgemein)	Kriterium (spezifisch)	Cohen's Kappa
Präaktionale Phase	Präaktionale Phase	0.75 (*)
	Vorbereitung/ Planung	0.29 (*)
	Emotionen/ Motivation	0.74 (*)
	Selbstwirksamkeit	0.19
	Ressourcen	0.55 (*)
	Einschätzung	0.31 (*)
Aktionale Phase	Aktionale Phase	0.71 (*)
	Aufgabenbearbeitung	0.33 (*)
	Lernstrategien	0.42 (*)
	Self-Monitoring	0.56 (*)
	Beobachtung	0.44 (*)
Post-Aktionale Phase	Post-Aktionale Phase	0.78 (*)
	Reflexion	0.55 (*)
	Konsequenzen	0.57 (*)
	Evaluation	0.38 (*)
Probleme	Zielbindung	0.45 (*)
	Self-Monitoring	0.49 (*)
	Regulation	-0.04
	Theorie/ Ansätze	0.51 (*)
	Teilkonstrukte	-0.03
	Zusatz	0.15 (*)

Die Überprüfung der Messgenauigkeit ergab bei den einzelnen Skalen auf allen Skalen eine niedrige Genauigkeit. Dies traf auf die Werte beider Hochschuldozenten zu (vgl. Tabelle 23). Die Gesamtskala ergab jedoch eine gute Messgenauigkeit (ebd., S. 129). Die Bewertungen sind bis auf die aktionale Phase, die Gesamtskala,- sowie die Probleme beim zweiten Hochschuldozenten nicht normalverteilt, d. h. es werden hierfür nicht-parametrische Verfahren für die Analyse herangezogen.

Tabelle 23 Skalenüberprüfung des selbstregulierten Lernens mithilfe des Cronbach's Alpha-Wertes

Kriterium (Allgemein) –Skalen	**α (1. Bewerter)**	**α (2. Bewerter)**	***Kolmogorov-Smirov-Test* (1. Bewerter)**	***Kolmogorov-Smirov-Test* (2. Bewerter)**
Präaktionale Phase	0.57	0.75	n. v.	n. v.
Aktionale Phase	0.56	0.70	normalverteilt	Normalverteilt
Post-Aktionale Phase	0.62	0.76	n. v.	n. v.
Probleme (ohne Zusatzpunkte)	0.02	0.46	-	-
Probleme (mit Zusatzpunkte)	-0.40	0.10	n. v.	Normalverteilt
Gesamtskala	0.82	0.87	normalverteilt	Normalverteilt

Die Untersuchung der Interkoderreliabilitäten (vgl. Tabelle 24) ergab bei allen Skalen eine niedrige Übereinstimmung (ebd., S. 129). Demzufolge bewerteten die beiden Begutachter die einzelnen Lernergebnisse auf Grundlage derselben Kriterien unterschiedlich.

Tabelle 24 Bewertungsübereinstimmung der einzelnen Skalen zum selbstregulierten Lernen (N = 103)

Kriterium (Allgemein)- Skalen	**Interkoderreliabilität**
Präaktionale Phase	0.75 (*)
Aktionale Phase	0.67 (*)
Post-Aktionale Phase	0.69 (*)
Probleme (mit Zusatzpunkte)	0.55 (*)
Gesamtskala	0.78 (*)

7.2.2.2 Lernstrategien

Die einzelnen Kriterien der Lernstrategien ergaben insgesamt eher eine niedrige Auswertungsobjektivität (vgl. Tabelle 25). Die Cohen's Kappa-Werte der Kategorien: Fragenstellen (Elaborationsstrategie), Schreiben (Wissensnutzungsstrategie) und Lernumgebung (Nutzung von Ressourcen) ergaben gute Übereinstimmungswerte zwischen den beiden Bewertern (Bühner, S. 129).

Tabelle 25 Bewertungsübereinstimmung der Kriterien zu den Lernstrategien

Lernstrategie (allgemein)	Lernstrategie (spezifisch)	Kriterium	Cohen's Kappa
Kognitive Lernstrategie	Elaboration	Vorwissen	0.21
		Fragen stellen	0.73 (*)
		Bilder	0.34 (*)
		Wiederholung	0.56 (*)
	Organisation	Wissensorganisation	0.09
		Textzusammenfassung	0.17
		Schemata	0.39 (*)
		Mind Maps	0.35 (*)
	Selbstregulation	Planung	0.58 (*)
		Überwachung	0.51 (*)
		Bewertung	0.34 (*)
		Regulation	0.34 (*)
	Wissensnutzung	Anwendung	0.12
		Problem lösen	0.38 (*)
		Schreiben	0.66 (*)
		Diskutieren	0.25
Motivation, Emotion		Indiv. Motivation	-0.03
		Lernumgebung	0.11
		Motiv.-emot. Bedingungen	0.46 (*)
Kooperatives Lernen		Auswirkung auf Motivation, Kognition	0.32 (*)
		Sozial-interaktive Lernformen	0.45 (*)
		Auswirkungen	0.49 (*)
Nutzung von Ressourcen		Nutzung von PC	0.18
		Digit. Lernressourcen	0.53 (*)
		Lernumgebung	0.72 (*)
		Zeit	0.18 (*)

Die Untersuchung der Reliabilitätswerte (Cronbach's Alpha) (vgl. Tabelle 26) ergab auf der Gesamtskala sowie auf den kognitiven Lernstrategien (Gesamt) eine gute Genauigkeit. Dies traf auf die Bewertungen beider Hochschuldozenten zu (vgl. Tabelle 26). Die Skalengenauigkeit war beim ersten Bewerter zudem bezüglich der Nutzung von Ressourcen-Strategie gut und beim zweiten Bewerter war sie bezüglich der Selbstregulation gut. Die Gesamtskala sowie alle kognitiven Strategien (Elaboration, Organisation, Selbstregulation, Wissensnutzung) sind beim ersten Hochschuldozent normalverteilt. Beim zweiten Hochschuldozenten trifft dies nur auf die Gesamtskala zu. Bei den nicht normalverteilten Bewertungen werden nicht-parametrische Verfahren für die Analyse herangezogen.

Tabelle 26 Skalenüberprüfung der Lernstrategien mithilfe des Cronbach's Alpha-Wertes

Lernstrategie (allgemein)	**Lernstrategie (spezifisch)**	**α (1. Bewerter)**	**α (2. Bewerter)**	***Kolmogorov-Smirnov-Test* (1. Bewerter)**	***Kolmogorov-Smirnov Test* (2. Bewerter)**
Lernstrategien Gesamt		0.84	0.89	normalverteilt	normalverteilt
Kognitive Lernstrategie	GESAMT	0.87	0.86	-	-
	Elaboration	0.58	0.56	normalverteilt	n. n
	Organisation	0.65	0.61	normalverteilt	n. n
	Selbstregulation	0.78	0.84	normalverteilt	n. n
	Wissensnutzung	0.71	0.71	n. n.	n. n
Motivation, Emotion		0.50	0.58	n. n.	n. n
Kooperatives Lernen		0.73	0.69	n. n.	n. n
Nutzung von Ressourcen		0.87	0.56	n. n.	n. n

n. n. = nicht normalverteilt

Die Untersuchung der Interkoderreliabilität (vgl. Tabelle 27) bei den einzelnen Skalen ergab eine gute Bewertungsübereinstimmung im Bereich der Elaboration, Organisation sowie Nutzung von Ressourcen. Alle anderen Skalen ergaben eine hohe Übereinstimmung zwischen den beiden Begutachtern (vgl. Tabelle 27).

Tabelle 27 Bewertungsübereinstimmung der einzelnen Skalen zu den Lernstrategien

Lernstrategie (allgemein)	Lernstrategie (spezifisch)	Interkoderreliabilität
Kognitive Lernstrategie	GESAMT	0.94
	Elaboration	0.86
	Organisation	0.89
	Selbstregulation	0.91
	Wissensnutzung	0.75
Motivation, Emotion		0.91
Kooperatives Lernen		0.93
Nutzung von Ressourcen		0.88

7.2.3 Hypothesenprüfende Darstellung

Zur Überprüfung der Hypothesen, ob sich die Bewertungen der Hochschuldozenten durch die Ähnlichkeitskennwerte abbilden lassen, werden zunächst Regressionsanalysen und anschließend die Korrelationen (nach Spearman) berechnet. Als Außenkriterien werden sowohl die Musterlösung, das Lernergesamtmodell und auch der Lehrtext herangezogen. Der Übersichtlichkeit wegen werden nur die signifikanten Regressionsanalysen tabellarisch dargestellt.

7.2.3.1 Selbstreguliertes Lernen

Die Regressionsanalysen zeigen, dass sich die inhaltlich orientierten Bewertungen beider Hochschuldozenten nicht durch die strukturellen Kennwerte vorhersagen lassen, was den theoretischen Vorannahmen entspricht. Die Bewertungen lassen sich durch die semantischen Kennwerte vorhersagen (vgl. Tabelle 28). Dies trifft auf das Gesamtmodell zu. Demzufolge lassen sich die vergebenen Leistungspunkte nicht auf Grundlage der kriterialen Bezugsnorm (Musterlösung und Lehrtext), sondern alleine auf Grundlage der sozialen Bezugsnorm vorhersagen. Dies trifft nicht auf das Gesamtmodell zu, in welchem die Stichpunktartigen Antworten von N = 18 Texte für die Analysen entfernt wurden.

Die erste aufgestellte Hypothese (H_1) darf in Bezug auf das Gesamtmodell (in welchem die Stichpunkte enthalten waren) angenommen werden – nicht jedoch in Bezug auf die Musterlösung und den Lehrtext.

Tabelle 28 Regressionsanalyse beider Bewerter

				GMM	
Kriterien	Kennwerte	df	Korr. R^2	ρ	F
Inhalt	Struktur	4	-0.03	0.87	0.31
Inhalt	Semantik	3	0.05	0.05	2.66

GMM= Gesamtmodell mit Stichpunkten

Die Untersuchung der Korrelationskoeffizienten (vgl. Tabelle 29) ergab bei den semantischen Kennwerten einen signifikanten Zusammenhang von $r = 0.23$ bezüglich des Concept Matching. Offensichtlich erhielten die Studierenden mehr Punkte in der Gesamtbewertung, wenn sie die Begrifflichkeiten in einen ähnlichen Zusammenhang wie dem im Gesamtmodell verwendeten.

Tabelle 29 Korrelationskoeffizienten beider Bewerter

	Kennwerte	**GM**
Struktur	Surface Matching	0.03
	Graphical Matching	0.00
	Structural Matching	0.06
	Gamma Matching	-0.08
Semantik	Concept Matching	0.23 (*)
	Propositional Matching	0.05
	Balanced Semantic Matching	-0.00

7.2.3.2 Lernstrategien

Die Regressionsanalysen zeigen, dass sich die inhaltlich orientierten Kriterien hinsichtlich des Lehrtextes durch die semantischen Kriterien abbilden lassen (vgl. Tabelle 30). Demzufolge lassen sich die Leistungspunkte mithilfe der kriterialen Bezugsnorm (Lehrtext als Referenzmodell) durch die semantischen Ähnlichkeitswerte vorhersagen. Dies trifft nicht auf die Musterlösung (kriteriale Bezugsnorm) und auch nicht auf das Gesamtmodell (soziale Bezugsnorm) zu. Die erste aufgestellte Hypothese (H_1) darf in Bezug auf den Lehrtext angenommen werden – nicht jedoch in Bezug auf die Musterlösung und das Gesamtmodell.

Tabelle 30 Regressionsanalyse beider Bewerter

				Lehrtext	
Kriterien	Kennwerte	df	Korr. R^2	ρ	F
Inhalt	Struktur	4	-0.03	0.63	0.65
Inhalt	Semantik	3	0.23	0.00	6.50

GMM war mit AKOVIA nicht analysierbar.

Die Überprüfung der Korrelationskoeffizienten (nach Spearman) (vgl. Tabelle 31) ergab bezüglich des Lehrtextes einen signifikanten Zusammenhang der semantischen Kennwerte Propositional Matching (r = 0.29) sowie Balanced Semantic Matching (r = 0.25).

Tabelle 31 Korrelationskoeffizienten beider Bewerter

	Kennwerte	Lehrtext
Struktur	Surface Matching	-0.11
	Graphical Matching	-0.03
	Structural Matching	-0.02
	Gamma Matching	0.08
Semantik	Concept Matching	0.21
	Propositional Matching	0.29 (*)
	Balanced Semantic Matching	0.25 (*)

7.2.4 Post-Hoc-Analyse

In einer Post-Hoc-Analyse wurden die Korrelationen zwischen den vergebenen Leistungspunkten und der Wortanzahl untersucht. Dies sollte absichern, dass die Hochschuldozenten die Lernergebnisse nicht gemäß der Textlänge bewerteten, d. h. dass längere Texte eine bessere Bewertung erzielten. Die deskriptive Betrachtung der durchschnittlichen Wortanzahl zeigte, dass die Texte bezüglich der Metakognition wesentlich kürzer ausfielen (vgl. Tabelle 32).

Tabelle 32 Deskription der Wortanzahl

	Selbstreguliertes Lernen (N = 103)				Lernstrategien (N = 70)			
	AM	SD	MD	Range	AM	SD	MD	Range
Wortanzahl	289.73	124.44	276	43- 784	207.14	136.13	183	14- 648

Die Analyse zeigte einen signifikanten Korrelationskoeffizienten (Spearman) betreffend der Textlänge sowie des Gesamteindrucks (vgl. Tabelle 33).

Tabelle 33 Wortanzahl und Leistungspunkte

	Selbstreguliertes Lernen (N = 103)	**Lernstrategien (N = 70)**
Bewerter	r	r
Dozent 1	0.60 (*)	0.58 (*)
Dozent 2	0.50 (*)	0.62 (*)

8 Ergebnisse der Schuluntersuchungen

Ausgehend von den in Kapitel sechs formulierten Fragestellungen und Hypothesen und der in Kapitel 5.3.7 ausgeführten methodischen Vorgehensweise, erfolgt die empirische Überprüfung. Diese umfasst eine deskriptive und daran anschließende hypothesenprüfende Darstellung der Ergebnisse. Daran anschließend folgen die Darstellungen der qualitativen Analysen sowie deren Reliabilitätsbestimmungen.

8.1 Ergebnisse der Untersuchungen im Fach Biologie

8.1.1 Bewertungskriterien Biologie

Die Analyse der (von den beiden Lehrkräften unabhängig voneinander entwickelten sowie angewendeten) Bewertungskriterien zeigte auf, dass sich diese unterschieden. Tabelle 34 bildet die Bewertungskriterien der ersten Lehrerkraft und Tabelle 35 die der zweiten Lehrkraft ab. Die von der ersten Lehrkraft angelegten Kriterien im Unterrichtsfach Biologie verdeutlichen ein nicht festgelegtes Spektrum an Antwortmöglichkeiten. Dabei umfassen sie keinen Katalog an inhaltlichen Aspekten, die notwendigerweise im Text enthalten sein mussten, um eine sehr gute Beurteilung zu erzielen. Ein Kriterium ist ausformuliert in Hinsicht auf den expliziten (inhaltlichen) Erwartungshorizont. Die anderen Kriterien umfassen Orientierungshilfen, die nicht explizit ausformuliert wurden.

Tabelle 34 Bewertungskriterien der ersten Lehrkraft

4 BE	Sprachlich, Gliederung, Materialnutzung, Menge
1 BE	Antwort
5 BE	Argumentation/ Begründung am Material
5 BE	Eigenes Material/ Beispiel
	Dabei müssen physische und neuronale Einflüsse angesprochen werden sowie positiv als auch negative Wirkungen in Beziehung gebracht werden.
∑ 15 BE	

Die Kriterien der zweiten Lehrkraft enthalten Schlagwörter und grenzen den Erwartungshorizont durch die einzeln aufgeführten und thematisch sortierten Begrifflichkeiten deutlich ein. Dabei ist genau festgelegt, für welchen Themenblock es wie viele Punkte gibt. Hier bleibt zunächst unklar, wie elaboriert diese in den Schülertexten ausformuliert sein müssen, um die angegebenen

Kriterien zu erfüllen. Aus Tabelle 35 wird deutlich, dass der Inhalt zwölf Punkte und der Kompetenzbereich drei Punkte von insgesamt fünfzehn Punkten ergab.

Tabelle 35 Bewertungskriterien der zweiten Lehrkraft

4 BE	Ecstasy illegale synthetische Droge als Pille oder Löschpapier – Trip, Amphetamine beabsichtigte Wirkung: z. B. Leistungssteigerung, Appetitzügler, Stimulans Wirkung im Körper: verstärkte Serotonin – Wirkung (Neurotransmitter) Nebenwirkungen: z. B. Herz – Kreislaufstörungen, Leberschäden
1 BE	LSD auf Basis Mutterkorn – löst Halluzinationen aus
3 BE	Alkohol legale Droge als Getränk beabsichtigte Wirkung: z. B: enthemmend und angstlösend Nebenwirkungen: z. B. Leberschäden, Organschäden
3 BE	Selbstgewähltes Beispiel z. B. Nikotin legale Droge im Tabak enthalten, durch Rauchen beabsichtigte Wirkung: z. B. beruhigend Nebenwirkungen: z. B. Lungenkrebsrisiko steigt
1 BE	Fazit Drogen – immer mit unerwünschte Nebenwirkungen, die zu gesundheitlichen Schäden führen (dosisabhängig)
3 BE	Kompetenzpunkte sinnvolle logische Gliederung richtiger Einsatz von Fachtermini und angemessener Ausdruck sinnvolle Verwendung der Materialien und Beschränkung auf das Wesentliche
∑ 15 BE	

Beide Lehrkräfte änderten ihre Bewertungskriterien nach dem Treatment nicht. Das Erstellen der textbasierten Musterlösung erfolgte durch die erste Lehrkraft. Anhang D zeigt den durch T - MITOCAR generierten Graph (vgl. Abbildung 13). Das Klassenmodell (Anhang D, Abbildung 14) erwies sich als wesentlich elaborierter als dies im Lehrermodell der Fall war. Die zweite Lehrkraft gab an, dass es unmöglich sei, einen idealen Erwartungshorizont explizit auszuformulieren. Da die textbasierte Musterlösung erst in der zweiten Bewertungsphase erstellt wurde, konnte hier keine Veränderung durch die Intervention festgestellt werden.

8.1.2 Ergebnisse

Die deskriptiven Analysen zeigen, dass die zweite Lehrkraft im Unterrichtsfach Biologie die Schülerergebnisse durchschnittlich höher einschätzte als die erste. Der Vergleich der Standardabweichungen sowie der Ränge zeigt, dass die Punktestreuung bei der zweiten Lehrkraft sich weiter ausbreitete als bei der ersten.

Die erste Lehrkraft vergab nach der Intervention niedrigere Punkte, wohingegen die zweite Lehrkraft die Schülerleistungen nach der Intervention durchschnittlich minimal als besser einschätzte. Insgesamt sind die durchschnittlichen Gesamteinschätzungen zwischen den beiden Messzeitpunkten eher stabil.

Tabelle 36 Deskription der Bewertungen im Pre-Post-Vergleich

	1 MP				**2 MP**			
	AM	SD	MD	Range	AM	SD	MD	Range
1. Lehrkraft	9.31	1.83	9	5-13	9.17	2.09	10	5- 12
2. Lehrkraft	10.86	2.31	11	6-15	10.97	2.31	11	5- 15

Die folgende Tabelle stellt die Ähnlichkeitswerte dar, welche die Nähe der Lernerergebnisse mit den Referenzmodellen abbildet. Die Referenzmodelle als Außenkriterium beinhalten zum einen die von der ersten Lehrkraft erstellte textbasierte Musterlösung und zum anderen alle Schülertexte (das Gesamtmodell der Lernergruppe). Die strukturellen Ähnlichkeiten der Lernergebnisse mit den Musterlösungen sind größer als die semantischen Ähnlichkeitskennwerte. Die hohen Werte bei Graphical Matching lassen vermuten, dass die Schüler ein ähnlich komplexes und konzeptuelles Wissen hatten wie das in den jeweiligen textbasierten Musterlösungen. Der hohe Gammawert deutet darauf hin, dass die einzelnen Schülertexte einen ähnlichen Grad an Vernetzung im Vergleich zu den jeweiligen Außenkriterien aufwiesen.

Tabelle 37 Deskription der Ähnlichkeitskennwerte (N = 29)

		ML		**GM**	
	Kennwerte	AM	SD	AM	SD
Struktur	Surface	0.51	0.23	0.37	0.22
	Graphical Matching	0.64	0.29	0.61	0.20
	Structural Matching	0.36	0.28	0.33	0.36
	Gamma	0.62	0.26	0.62	0.25
Semantik	Concept Matching	0.22	0.11	0.33	0.12
	Propositional Matching	0.04	0.06	0.16	0.11
	Balanced Semantic Matching	0.13	0.19	0.45	0.23

ML = Musterlösung; GM = Gesamtmodell beider Klassen

8.1.3 Qualität der Bewertungskriterien und Bewertungsstabilität

Die Qualität der Bewertungskriterien wurde mithilfe des Alphawertes bestimmt. Diese sind bei der Gesamteinschätzung (N = 3 Inhalt, Sprache, Gesamt) gut.

Tabelle 38 Skalenüberprüfung mithilfe des Cronbach's Alpha-Wertes

	1 MP α	**2 MP** α	**1 MP** Kolmogorov-Smirov-Test Normalverteilung	**2MP** Kolmogorov-Smirov-Test Normalverteilung
1. Lehrkraft – Inhalt			normalverteilt	Normalverteilt
1. Lehrkraft - Sprache			n. n.	Normalverteilt
1. Lehrkraft - Gesamt	0.85	0.82	normalverteilt	Normalverteilt
2. Lehrkraft - Inhalt			normalverteilt	Normalverteilt
2. Lehrkraft - Struktur			n. n.	n. n.
2. Lehrkraft - Gesamt	0.81	0.83	normalverteilt	Normalverteilt

Die Untersuchung der Interkoderreliabilitäten vor und nach dem Treatment ist in den folgenden Tabellen abgebildet. Tabelle 39 zeigt die Korrelationskoeffizienten (Spearman) zwischen den von den Lehrkräften vergebenen Leistungspunkten (N = 29 Schüler) vor und nach der Schulung. Sie verdeutlichen, dass sich nur die strukturell orientierten Bewertungen zwischen den beiden Lehrkräften nicht ähnelten. Beide Lehrkräfte schätzten die Schülertexte für die strukturell orientierten Kriterien unterschiedlich ein. Die Gesamteinschätzung sowie die inhaltlich orientierten Bewertungen ähnelten sich gering (vgl. Bühner, 2004, S.129). Dies lässt sich durch die unterschiedlich zugrunde gelegten Kriterien begründen.

Tabelle 39 Interkoderreliabilität zwischen den Lehrkräften

		r
1 MP	Bewertung insgesamt	0.48 (*)
	Bewertung (inhaltlich)	0.44 (*)
	Bewertung (strukturell)	0.20
2 MP	Bewertung insgesamt	0.50 (*)
	Bewertung (inhaltlich)	0.60 (*)
	Bewertung (strukturell)	0.11

Die Untersuchung der Interkoderreliabilitäten ergab bei wiederholter Bewertung eine hohe Bewertungsübereinstimmung bezüglich der zweiten Lehrkraft. Sie waren bei der Gesamtbewertung sowie der inhaltlich orientierten Bewertung zuverlässig. Die Stabilität der strukturell orientierten Kriterien war niedrig. Bei der ersten Lehrkraft ergaben sich bei wiederholter Bewertung niedrige Übereinstimmungen

bei allen Einschätzungen (Gesamtbewertung, inhaltlich orientierte Bewertung sowie strukturell orientierte Bewertung). Demzufolge genügte die strukturell orientierte Bewertung auf Grundlage beider Kriterienkatalogen nicht den wissenschaftlichen Standards bezüglich der Zuverlässigkeit. Die Einschätzungen der zweiten Lehrkraft deuten auf eine hohe reliable Messung der Gesamteinschätzungen sowie der inhaltlichen Bewertung hin.

Tabelle 40 Interkoderreliabilität innerhalb der Lehrkräfte

		r
Lehrer 1	Bewertung insgesamt	0.59 (*)
	Bewertung (inhaltlich)	0.66 (*)
	Bewertung (strukturell)	0.40 (*)
Lehrer 2	Bewertung insgesamt	0.91 (*)
	Bewertung (inhaltlich))	0.92 (*)
	Bewertung (strukturell)	0.79 (*)

8.1.4 Hypothesenprüfende Darstellung

Der folgende Teil umfasst die Überprüfung der im Kapitel fünf aufgestellten statistischen Hypothesen. Die Überprüfung inwiefern die beiden Lehrkräfte auf Grundlage unterschiedlicher Kriterienkataloge (vgl. Tabelle 34 und 35) zu ähnlichen Gesamteinschätzungen gelangen, ist bereits im vorangegangenen Abschnitt (8.1.3) dargestellt.

Die Analysen mithilfe der Regressionsanalyse zeigen, dass sich die Bewertungen beider Lehrkräfte (sowohl die inhaltlich orientierten als auch die Layout orientierten) nicht durch die Kennwerte (weder durch die semantischen noch durch die strukturellen) abbilden lassen (siehe Anhang C, Tabellen 109 - 112). Weiter zeigt das korrigierte R^2, dass die Leistungsermittlung beider Lehrkräfte nicht durch die Ähnlichkeitsmaße erklärt werden kann. Deswegen dürfen die signifikanten Korrelationskoeffizienten nicht interpretiert werden, denn es kann nicht ausgeschlossen werden, dass diese zufällig zustande gekommen sind. Die aufgestellten Hypothesen (H_2 und H_3) dürfen nicht angenommen werden.

8.1.5 Post-Hoc-Analyse

Die deskriptive Betrachtung ergab eine durchschnittliche Wortanzahl von 346 Wörtern. Der kürzeste Schülertext umfasste 233 Wörter und der längste Schülertext 515 Wörter (vgl. Tabelle 41).

Tabelle 41 Deskription der Wortanzahl (N = 29)

	AM	SD	MD	Range
Wortanzahl	346	61.31	343	233- 515

Eine Post-Hoc-Analyse zeigte bezüglich der Gesamteinschätzung sowie der inhaltlich orientierten Einschätzungen der ersten Lehrkraft einen signifikanten Korrelationskoeffizienten (Spearman). Es kann somit nicht ausgeschlossen werden, dass die Textlänge die Bewertungen der ersten Lehrkraft beeinflussten. Bei der zweiten Lehrkraft hatte die Textlänge keinen Einfluss auf die Bewertungen (vgl. Tabelle 42).

Tabelle 42 Wortanzahl und Leistungspunkte

		1 MP	2 MP
Lehrkraft 1	Gesamteinschätzung	0.48 (*)	0.39 (*)
	Inhaltlich orientierte Einschätzung	0.50 (*)	0.42 (*)
	Strukturell orientierte Einschätzung	0.24	0.09
Lehrkraft 2	Gesamteinschätzung	0.21	0.27
	Inhaltlich orientierte Einschätzung	0.19	0.27
	Strukturell orientierte Einschätzung	0.18	0.18

MP = Messzeitpunkt

8.1.5.1.1 Güte der Musterlösung

Zum Schluss wurde überprüft, inwiefern die in den Kriterien enthaltenen Aspekte in der textbasierten Musterlösung enthalten sind. Die Befunde zeigen, dass die (von der ersten Lehrkraft) erstelle Musterlösung aus teilweise nicht ausgeschriebenen Wörtern bestand. Eine Gliederung in: Einleitung, Hauptteil, Schluss war schwer zu erkennen. Die von der ersten Lehrkraft angegebenen inhaltlichen Kriterien waren in der Musterlösung abgebildet. Bei der zweiten Lehrkraft, die in dieser Untersuchung keine textbasierte Musterlösung erstellte, ergaben die Befunde bei den strukturellen Kriterien (Kompetenzpunkte: richtiger Einsatz von Fachtermin), dass diese vorhanden waren, die Fachbegriffe jedoch ungenau erklärt wurden. Die Beschränkung auf das Wesentliche sowie die Verwendung der Materialien war gegeben. Die inhaltlichen Kriterien der zweiten Lehrkraft zeigten, dass das Kriterium des selbst gewählten Beispiels ganz und das Kriterium Ecstasy teilweise vorhanden war. LSD und Alkohol waren nicht in der Musterlösung abgebildet (vgl. Tabelle 43).

Tabelle 43 Güte der Musterlösung

	Kriterium		
Lehrkraft 1	Struktur	Sprachlich	Wörter nicht ausgeschrieben
		Gliederung	schwer erkennbar
	Inhalt		Vorhanden
Lehrkraft 2	Struktur	Gliederung	schwer erkennbar
		richtiger Einsatz von Fachtermini	Ja, allerdings unpräzise Erklärung
		Ausdruck	viele Abkürzungen
		Verwendung der Materialien und Beschränkung auf das Wesentliche	Vorhanden
	Inhalt	Ecstasy	teilweise vorhanden
		Selbstgewähltes Bsp.	Vorhanden
		LSD + Alkohol	nicht in Musterlösung enthalten
		Fazit	nicht in Musterlösung vorhanden

8.2 Ergebnisse der Untersuchungen im Fach Deutsch

8.2.1 Bewertungskriterien

Die Bewertungskriterien zeigen, dass eine relativ offene Bewertung vorliegt. Es sind keine Schlagwörter oder inhaltlichen Vorgaben enthalten, die in der Erörterung gegeben sein mussten. Vielmehr spezifizieren die einzelnen inhaltlich orientierten Kriterien das formale Kriterium einer erkennbaren dreigliedrigen Struktur näher. Vor der Intervention erfolgte eine Einteilung des Vorhandenseins der einzelnen Merkmale in den einzelnen Schülertexten in „++“, „+“, „0“, „-“ sowie „--“. Nach der Intervention schätzte die Lehrkraft das Vorhandensein der einzelnen Kategorien in den einzelnen Schülertexten auf einer Punkteskala von eins bis fünf ein, um daraus eine mögliche Tendenz hinsichtlich der Notengebung zu bekommen. Im Interview verwies sie darauf, dass sie bei der zweiten Bewertungsrunde den einzelnen Kategorien Punkte vergeben hatte, um daraus genau die Note festzulegen. Vor der Intervention, so gab die Lehrkraft an, hatte sie die Kriterien eher als grobe Einschätzung angesehen, um den Schülern auf deren Grundlage eine Rückmeldung zu geben in welche Richtung deren Leistungen gingen. Nach der Intervention vergab die Lehrkraft je nach Ausprägungsgrad jeweils 0.5 Punkte und verrechnete diese später zu einer Gesamtnote. Das Erstellen der textbasierten Musterlösung erfolgte in der zweiten Bewertungsrunde. Deswegen konnte hier keine Veränderung durch die Schulung festgestellt werden.

Tabelle 44 Bewertungskriterien im Unterrichtsfach Deutsch

Kriterium					
Vorher	--	-	0	+	++
Nachher	1	2	3	4	5
Inhalt					
Wird in der Einleitung das Thema deutlich benannt?					
Gibt es einen Überleitungssatz zum Hauptteil?					
Wurden die Argrumente sinnvoll angeordnet? (von schwach zu stark)					
Wurden die Argumente durch Beispiele gestützt?					
Wurden die Argumente sinnvoll ausgewählt?					
Enthält der Schluss einen Ausblick/ eine persönliche Meinung/ eine weiterführende Fragestellung?					
Sprache					
Wurden die Argumente mit unterschiedlichen Überleitungen eingeleitet?					
Ist der sprachliche Ausdruck angemessen?					
Sind Rechtschreibung, Grammatik und Zeichensetzung weitgehend in Ordnung?					
Form					
Ist eine dreigliedrige Struktur eindeutig erkennbar (Einleitung, Hauptteil, Schluss)?					

8.2.2 Ergebnisse

Die deskriptiven Analysen zeigen, dass die Lehrkraft im Unterrichtsfach Deutsch die Schülerergebnisse nach der Intervention durchschnittlich niedriger einschätzte. Der Vergleich der Standardabweichungen sowie der Ränge zeigt, dass die Notenvergabe vor der Intervention höher streute. Insgesamt veränderte sich der durchschnittliche Gesamteindruck der Lehrkraft von 2.26 auf 2.56 Notenpunkte.

Tabelle 45 Deskription der Bewertungen

1 MP				**2 MP**			
AM	SD	MD	Range	AM	SD	MD	Range
2.26	0.65	2.25	1.25-3.50	2.56	0.50	2.50	1.25- 3.25

MP = Messzeitpunkt, MD = Median; 1 = sehr gut, 2 = gut, 3 = befriedigend, 4 = ausreichend, 5 = mangelhaft, 6 = ungenügend

Die folgende Tabelle stellt die strukturellen und semantischen Übereinstimmungen der Schülerleistungen mit den Referenzmodellen dar. Die durchschnittlichen strukturellen Ähnlichkeiten der einzelnen Lernergebnisse sind mit beiden Außenkriterien größer als die durchschnittliche semantische Übereinstimmung.

Dies weist darauf hin, dass die einzelnen Schülertexte ein durchschnittlich ähnlich breites sowie komplexes Wissen externalisierten wie das der Lehrkraft.

Tabelle 46 Deskription der Ähnlichkeitskennwerte

Kennwerte		ML		GM	
		AM	SD	AM	SD
Struktur	Surface Matching	0.62	0.16	0.62	0.16
	Graphical Matching	0.78	0.18	0.78	0.18
	Structural Matching	0.55	0.22	0.55	0.22
	Gamma Matching	0.82	0.15	0.82	0.15
Semantik	Concept Matching	0.37	0.05	0.37	0.05
	Propositional Matching	0.15	0.11	0.15	0.11
	Balanced Semantic Matching	0.40	0.27	0.40	0.27

ML = Musterlösung; GM = Gesamtmodell

Die Musterlösungen waren sich auf der graphentheoretischen Ebene ähnlich genug, um gleiche durchschnittliche Ähnlichkeitskennwerte mit den einzelnen Lernermodellen zu identifizieren. Das zeigt, dass das Gesamtmodell (welches als soziale Bezugsnorm betrachtet werden darf), ein ähnlich breites Wissen abbildete, wie das der Lehrkraft und dass darin ähnliche Begrifflichkeiten abgebildet waren wie die der Lehrkraft. Diese Ähnlichkeit kann auch mit den T-MITOCAR generierten Graphen erklärt werden (vgl. Anhang D, Abbildung 15 und 16). Diese verdeutlichen, dass die Schüler ähnliche Begrifflichkeiten wie die der Lehrkraft verwenden. Ebenso ist eine ähnliche Dichte der Vernetztheit erkennbar.

8.2.3 Bewertungsstabilität

Die Analyse der Einschätzungsstabilität zwischen den einzelnen Messzeitpunkten ergab einen signifikanten Korrelationskoeffizienten (Spearman) von r = 0.63 (bei N = 17 Lernenden). Dies deutet auf eine niedrige Bewertungsübereinstimmung hin (vgl. Bühner, 2004, S. 129). Demzufolge liegt bezüglich der Bewertungen eine unzureichende Messzuverlässigkeit vor.

Tabelle 47 Interkoderreliabilität innerhalb der Lehrkraft

	r
Bewertung insgesamt	0.63 (*)

8.2.4 Hypothesenprüfende Darstellung

Die Regressionsanalysen zeigten, dass die strukturellen Kennwerte beim ersten Messzeitpunkt mit der Gesamteinschätzung signifikant korrelierten. Dies lässt sich durch die Kriterien erklären, die primär strukturell orientierte Aspekte für die Anordnung des Textes umfassten. Beim zweiten Messzeitpunkt ergab sich zwischen den strukturellen Kennwerten und den formal orientierten Kriterien keine Regression. Das lässt darauf schließen, dass die formal orientierten Kriterien etwas anderes abbilden als die strukturellen Ähnlichkeitskennwerte. Auch zwischen den inhaltlich orientierten Kriterien und den semantischen Ähnlichkeitskennwerten konnte kein bedeutsamer Zusammenhang festgestellt werden. Die Hypothese (H_5) - nicht jedoch H_4 - darf beim ersten Messzeitpunkt angenommen werden. Es besteht ein Zusammenhang zwischen den Gesamtbewertungen und den strukturellen Kennwerten.

Tabelle 48 Regressionsanalyse des ersten Messzeitpunkts (Gesamtbewertung) (N = 17)

		Lehrermodell			Klassenmodell		
Kennwerte	df	Korr. R^2	ρ	F	Korr. R^2	ρ	F
Struktur	4	0.72	0.00	11.30	0.72	0.00	11.30
Semantik	3	-0.17	0.86	0.25	-0.17	0.86	0.25

Beim ersten Messzeitpunkt zeigte sich bezüglich des strukturellen Kennwertes Gamma ein bedeutsamer Zusammenhang von $r = -0.42$ in Bezug auf die Gesamteinschätzung. Dies lässt darauf deuten, dass wenn die einzelnen Schülertexte Begrifflichkeiten ähnlich elaboriert verwendeten wie die in den jeweiligen Außenkriterien (textbasierte Musterlösung der Lehrkraft sowie das Gesamtmodell aller Schüler), sie eine bessere Note erhielten. Die Analysen ergaben, dass die Korrelationswerte der Lehrerlösung identisch mit denen des Klassengesamtmodells waren.

Tabelle 49 Korrelationskoeffizienten (Lehrerlösung und Gesamtmodell)

		1 MP
	Kennwerte	Gesamtbewertung
Struktur	Surface Matching	-0.39
	Graphical Matching	-0.38
	Structural Matching	-0.04
	Gamma Matching	-0.42 (*)
Semantik	Concept Matching	0.13
	Propositional Matching	0.14
	Balanced Semantic Matching	0.16

8.2.5 Post-Hoc-Analyse

Die deskriptive Untersuchung zeigt eine durchschnittliche Wortanzahl von 679 Wörtern. Die kürzeste Erörterung umfasste 320 Wörter und die längste Erörterung 1179 Wörter (vgl. Tabelle 50).

Tabelle 50 Deskription der Wortanzahl

	AM	SD	MD	Range
Wortanzahl	678.65	244.81	615	320- 1179

Die Betrachtung der Gesamteinschätzung und der Wortanzahl ergab keinen Zusammenhang (vgl. Tabelle 51). Dies deutet darauf hin, dass die Lehrkraft sich in der Bewertung nicht durch die Textlänge beeinflussen ließ.

Tabelle 51 Wortanzahl und Gesamteindruck

	1 MP	2 MP
Gesamtbewertung	-0.29	-0.27

8.2.6 Güte der Musterlösung

Die Untersuchung, inwiefern alle in den Kriterien abgebildeten Aspekte in der textbasierten Musterlösung enthalten waren, ergab folgendes. Bei den inhaltlichen Kriterien war der Überleitungssatz zum Hauptteil nicht vorhanden. Die Kriterien der sinnvollen Anordnung der Argumente sowie der beispielhaften Unterstützung waren teilweise vorhanden. Alle anderen inhaltlichen Kriterien waren in der Musterlösung enthalten. In Hinsicht auf die sprachlichen sowie formalen Kriterien waren alle Kriterien vorhanden (vgl. Tabelle 52).

Tabelle 52 Güte der Musterlösung

Kriterium	Vorhanden	teilweise vorhanden	nicht vorhanden
Inhalt			
Wird in der Einleitung das Thema deutlich benannt?	X		
Gibt es einen Überleitungssatz zum Hauptteil?			X
Wurden die Argrumente sinnvoll angeordnet? (von schwach zu stark)		X	
Wurden die Argumente durch Beispiele gestützt?		X	
Wurden die Argumente sinnvoll ausgewählt?	X		
Enthält der Schluss einen Ausblick/ eine persönliche Meinung/ eine weiterführende Fragestellung?	X		
Sprache			
Wurden die Argumente mit unterschiedlichen Überleitungen eingeleitet?	X		
Ist der sprachliche Ausdruck angemessen?	X		
Sind Rechtschreibung, Grammatik und Zeichensetzung weitgehend in Ordnung?	X		
Form			
Ist eine dreigliedrige Struktur eindeutig erkennbar (Einleitung, Hauptteil, Schluss)?	X		

8.3 Ergebnisse der ersten Teilstudie im Fach Religion

8.3.1 Bewertungskriterien

Folgende Tabellen (53-55) stellen die (von den beiden Lehrkräften gemeinsam entwickelten) Bewertungskriterien dar. Diese umfassen zunächst eine generelle Bewertung in Bezug auf sprachliche Aspekte (wie z. B. Ausdruck, Grammatik, Rechtschreibung) sowie auf die einzelnen Aufgabenstellungen hin bezogenen inhaltlichen Kriterien. Aus der Summe wird deutlich, dass bei der generellen Bewertung keine Punkte vergeben wurden. Beide Lehrkräfte veränderten ihre Kriterien und ihre textbasierte Musterlösung nach der Schulung nicht. Die zweite Lehrkraft gab an, die Gewichtungen für den Inhalt sowie die Struktur nach der Intervention verändert zu haben. Diese veränderten sich von der ursprünglichen Verteilung: 70% auf Inhalt und 30 % auf Struktur zu 80% auf Inhalt und 20% auf Struktur. Die Musterlösungen unterschieden sich minimal zueinander. Die Ähnlichkeit der Musterlösung ist auch in den durch T-MITOCAR generierten Graphen (vgl. Anhang D, Abbildung 17 und 19) erkennbar. Dies wurde von Seiten der Lehrkräfte darin begründet, dass berücksichtigt wurde, was in den jeweiligen Klassen behandelt worden war und was nicht. Die Kriterien unterschieden sich jedoch nicht voneinander. Dabei zeigen die inhaltlichen Kriterien bei der ersten Aufgabenstellung Aspekte, die die Organisation des Textes umfassen (wie beispielsweise das Kennzeichnen von Zitaten sowie die Verwendung des Konjunktives).

Tabelle 53 Bewertungskriterien im Unterrichtsfach Religion

			Punkte
Generelle Bewertung	Ausdruck/ Sprachliche Leistung	Satzbau, Art der Formulierung, Einbezug der Fachbegriffe Bei entsprechenden Mängeln bis zu 1 Punkt Abzug	
	Grammatik, Rechtschreibung und Zeichensetzung	Einhaltung des orthographischen Regelkatalogs Bei Häufung der Fehler bis zu 2 Punkte Abzug	
Aufgabe 1	Nennung von Autor, Quelle und Schwerpunkt des Textes	Franz Alt, Auszug aus „Frieden ist möglich“ (1986) z. B. Auslegung/ Umsetzung der von Jesus geforderten Feindesliebe	0.5
	Zusammenfassung der Hauptthesen	- Feindesliebe ist Alternative zur Kriegspolitik (0,5 P) - Wichtig ist, dass man ohne Einschränkung („Vorbehalt“) friedliebend sein will (1 P) - Am Frieden muss man durch „Selbstverwirklichung“ arbeiten (1 P) - Wahre Selbstverwirklichung und Menschlichkeit solidarisiert sich mit anderen; als Antwort auf die Menschwerdung Gottes (1 P) - Dadurch ist „Heilung der Welt“ möglich (1 P)	4.5
	Verwendung von eigenen Worten	Eigene Worte/ zu nah am Text („Zitatencollage“)	
	Kennzeichnung von Zitaten	Erfolgt konsequent/ überwiegend/ teilweise/ nicht	0.5
	Konjunktiv	Konsequenter/ überwiegender/ teilweiser/ kein Gebrauch	0.5
		Summe	6.0

Tabelle 54 Bewertungskriterien im Unterrichtsfach Religion

Aufgabe 2	Jüdische Erwartung	- Messias als zukünftiger gerechter Herrscher, der die sozialen Verhältnisse radikal ändert (1 P.) - Befreiung von der römischen Besatzungsmacht (1 P.) - Zeloten wollten das Reich Gottes mit Gewalt schneller herbeizwingen (1 P.)	3.0
	Erwartung von Jesus	- Reich Gottes schafft Neues: Vergebung, Heilung, Speisung, Todesüberwindung (Textbezug: Z. 73; mögliche Gleichnisse zur Verdeutlichung: „Verlorener Sohn“, „Arbeiter im Weinberg“) (1 P.) - Gilt besonders für Außenseiter, Traurige, Leidende, Arme, Kinder, Frauen (1 P.) - durch Gott und unter Mitwirkung des Menschen, obgleich ohne Gewalt (Textbezug: Z.21/22) (1 P.) - Reich Gottes hat bereits zeichenhaft angefangen und wird noch vollendet => „eschatologischer Vorbehalt“ (Textbezug: Z.19-21/23/81; mögliches Gleichnis zur Verdeutlichung: „Gleichnis vom Senfkorn“) (1 P.)	4.0
	Fazit/Vergleich	Jesu Vorstellung vom Reich Gottes widerspricht der jüdischen Erwartungshoffnung: keine radikale Umwälzung sondern zeichenhafte Veränderung, die bei dem Einzelnen anfängt (1 P.)	1.0
		Summe	8.0

Tabelle 55 Bewertungskriterien im Unterrichtsfach Religion

	Mögliche Ansätze	
Aufgabe 3	Bergpredigt ist für Menschen mit besonderer Heiligkeit (z. B. Mönche, Asketen)	Z. B. Martin Luther King, Gandhi mit gewaltlosem Widerstand
	Interimsethik für die Zeit von Jesus und seinen Jüngern	Erwartetes Weltende ist nicht gekommen, deswegen hat diese Ethik keine unmittelbare Bedeutung mehr, somit auch nicht die Feindesliebe, die in der erwarteten kurzen Zeit bis zum Weltende hätte erduldet werden können
	Erzeugung einer moralischen Gesinnung und eines guten Willens	Bergpredigt nicht als grundsätzlicher Regelkatalog zu verstehen sondern als Inspiration -> keine generelle Feindesliebe sondern in Einzelfällen
	Unterscheidung zwischen dem Menschen als Christen und als Amtsperson (z. B. Richter, Soldat)	Feindesliebe gilt nur für einzelne, politisch lässt sich danach nicht handeln
	Bergpredigt als bewusste Überforderung und Unmöglichkeit	Menschen können Feinde nicht lieben -> Spiegel, in dem man die eigene Sünde erkennt und dadurch auf die Gnade Gottes angewiesen ist
	Eine Lesart der Feindesliebe nach Franz Alt	
	Abschlussfazit	

Bewertung bei Aufgabe 3 je nach Aufbau, Argumentation, Sprache und Inhalt individuell = 6 Punkte

8.3.2 Ergebnisse

Die deskriptiven Analysen zeigen, dass die erste Lehrkraft im Unterrichtsfach Religion die Schülerergebnisse nach der Intervention durchschnittlich niedriger einschätzte. Der Vergleich der Standardabweichungen sowie der Ränge zeigt, dass die Punkte vor der Intervention weiter streuten als nach der Intervention.

Tabelle 56 Deskription der Bewertungen der ersten Lehrkraft

1 MP				**2 MP**			
AM	SD	MD	Range	AM	SD	MD	Range
11.67	1.80	12	8- 14	10.20	1.52	10	8- 12

Die folgende Tabelle stellt die Ähnlichkeitswerte dar, welche die Nähe der Lernerergebnisse mit den Referenzmodellen abbildet. Der hohe Concept Matching Wert zeigt, dass die Schüler Begrifflichkeiten ähnlich verwendeten wie die

Lehrkraft. Der hohe Balanced Semantic Wert weist darauf hin, dass wenn die Schüler ähnliche Begrifflichkeiten wie die der Lehrkraft verwendeten, diese in einen ähnlichen Zusammenhang gebracht wurden. Dies trifft auch auf das Gesamtmodell zu, die Schüler verwendeten im Modell übereinstimmende Begrifflichkeiten im selben Zusammenhang.

Tabelle 57 Deskription der Ähnlichkeitskennwerte (N = 15)

	Kennwerte	**ML**		**GM**	
		AM	SD	AM	SD
Struktur	Surface Matching	0.69	0.21	0.58	0.23
	Graphical Matching	0.77	0.16	0.69	0.19
	Structural Matching	0.61	0.19	0.53	0.31
	Gamma Matching	0.79	0.14	0.68	0.16
Semantik	Concept Matching	0.54	0.11	0.55	0.09
	Propositional Matching	0.30	0.12	0.32	0.12
	Balanced Semantic Matching	0.56	0.18	0.58	0.19

ML = Musterlösung; GM = Gesamtmodell

8.3.3 Bewertungsstabilität

Die Untersuchung der Einschätzungsstabilität ergab einen signifikanten Korrelationskoeffizienten (Spearman) von r = 0.64 (vgl. Tabelle 58). Dies lässt eine instabile Bewertung vermuten (vgl. Bühner, 2004, S.129). Demnach genügt die vorliegende Messzuverlässigkeit nicht den wissenschaftlichen Standards.

Tabelle 58 Interkoderreliabilität zwischen den Messzeitpunkten

	r
Bewertung insgesamt	0.64 (*)

MP = Messzeitpunkt

8.3.4 Hypothesenprüfende Darstellung

Die Regressionsanalysen ergaben, dass sich die Gesamtbewertung der ersten Lehrkraft weder durch die strukturellen noch durch die semantischen Kennwerte abbilden lässt. Dies zeigt, dass die Kriterien etwas anderes abbildeten als die Kennwerte. Die Hypothesen (H_4 und H_5) dürfen nicht angenommen werden.

8.3.5 Post-Hoc-Analyse

Eine Post-Hoc-Untersuchung zeigte, dass die Schülertexte durchschnittlich 649 Wörter beinhalteten. Die kürzeste textbasierte Antwort auf die drei gestellten Aufgabenstellungen umfasste 363 Wörter und die längste Antwort 869 Wörter (vgl. Tabelle 59).

Tabelle 59 Deskription der Wortanzahl

	AM	SD	MD	Range
Wortanzahl	649.07	127.91	662	363- 869

Die Analyse der Gesamtbewertung im Zusammenhang mit der Textlänge ergab beim ersten Messzeitpunkt einen signifikanten Korrelationskoeffizienten von $r = 0.77$ und beim zweiten Messzeitpunkt einen signifikanten Korrelationskoeffizienten von $r = 0.54$ (vgl. Tabelle 60). Dies deutete darauf hin, dass sich die Lehrkraft bei der Bewertung von der Textlänge beeinflussen ließ. Es kann jedoch nicht ausgeschlossen werden, dass die längeren Texte auch wirklich qualitativ besser waren. Der Einfluss der Textlänge nahm nach der Intervention deutlich ab und lässt vermuten, dass die Lehrkraft bei der zweiten Bewertung bewusst darauf achtete, sich nicht von der Textlänge beeinflussen zu lassen. Die Minimierung dieses Fehlers war Inhalt der Schulung gewesen.

Tabelle 60 Wortanzahl und Gesamteindruck

	1 MP	2 MP
Gesamtbewertung	0.77 (*)	0.54 (*)

8.3.6 Güte der Musterlösung

Die Untersuchung, inwiefern die in den Kriterien enthaltenen Aspekte in der textbasierten Musterlösung enthalten waren, ergab folgendes. Bei der ersten Aufgabenstellung beinhaltete die Musterlösung alle Kriterien bis auf das Kriterium in Hinsicht auf die Verwendung von eigenen Worten. Dieses war nur teilweise vorhanden. Bei der zweiten Aufgabenstellung waren alle Kriterien in der Musterlösung enthalten bis auf das Kriterium hinsichtlich der Erwartung von Jesus, welches teilweise vorlag (vgl. Tabelle 61). Aufgabe drei war nicht in der Musterlösung abgebildet.

Tabelle 61 Güte der Musterlösung

	Kriterium	vorhanden	teilweise vorhanden	nicht vorhanden
Aufgabe 1	Nennung von Autor, Quelle und Schwerpunkt des Textes	X		
	Zusammenfassung der Hauptthesen	X		
	Verwendung von eigenen Worten		X	
	Kennzeichnung von Zitaten	X		
	Konjunktiv	X		
Aufgabe 2	Jüdische Erwartung	X		
	Erwartung von Jesus		X	
	Fazit/ Vergleich	X		

8.4 Ergebnisse der zweiten Teilstudie im Fach Religion

8.4.1 Ergebnisse

Die Betrachtung der Mittelwerte zeigte, dass die zweite Lehrkraft die Schülerergebnisse nach der Intervention durchschnittlich höher einschätzte als vor der Intervention. Der Vergleich der Standardabweichungen macht deutlich, dass die Punkte nach der Intervention mehr streuten als vor der Intervention.

Tabelle 62 Deskription der Bewertungen der zweiten Lehrkraft

1 MP				2 MP			
AM	SD	MD	Range	AM	SD	MD	Range
15.75	4.97	16.50	8- 25	16.25	5.58	16.00	9- 26

MP = Messzeitpunkt

Die Betrachtung der Ähnlichkeitswerte zeigt, dass die strukturellen Ähnlichkeiten der Lernergebnisse mit den Musterlösungen größer als die semantischen Ähnlichkeitskennwerte sind (vgl. Tabelle 63). Die hohen Graphical Matching- und Gamma-Werte verdeutlichen, dass die Schülermodelle ein ähnlich breites Wissen abbildeten wie das Lehrermodell; während die hohen Gammawerte auf eine ähnliche Knotendichte der Schülermodelle hinsichtlich der Referenzmodelle hindeuten.

Tabelle 63 Deskription der Ähnlichkeitskennwerte (N = 12)

Kennwerte		ML		GM	
		AM	SD	AM	SD
Struktur	Surface Matching	0.59	0.14	0.59	0.25
	Graphical Matching	0.75	0.21	0.62	0.18
	Structural Matching	0.64	0.22	0.63	0.31
	Gamma Matching	0.61	0.17	0.73	0.15
Semantik	Concept Matching	0.36	0.16	0.53	0.13
	Propositional Matching	0.19	0.15	0.30	0.14
	Balanced Semantic Matching	0.43	0.29	0.55	0.19

ML = Musterlösung; GM = Gesamtmodell

8.4.2 Bewertungsstabilität

Die Überprüfung der Einschätzungsstabilität der zweiten Lehrkraft ergab einen signifikanten Korrelationskoeffizienten (Spearman) von $r = 0.75$ (bei N = 12 Schüler). Dies deutet auf eine geringe Bewertungsübereinstimmung zwischen den Messzeitpunkten hin (vgl. Bühner, 2004, S.129). Demzufolge genügte die Messzuverlässigkeit nicht um den wissenschaftlichen Standards zu genügen.

Tabelle 64 Interkoderreliabilität zwischen den Messzeitpunkten

	r
Bewertung insgesamt	0.75 (*)

8.4.3 Hypothesenprüfende Darstellung

Die Regressionsanalysen zeigten, dass sich die Gesamtbewertungen der Lehrkraft beim ersten Messzeitpunkt hinsichtlich des Gesamtmodells (als soziale Bezugsnorm) durch die strukturellen Kennwerte abbilden lassen. Die Hypothese (H_5) darf angenommen werden. Es besteht ein Zusammenhang zwischen der Gesamtbewertung und den strukturellen Kennwerten – jedoch nicht zwischen der Gesamtbewertung und den inhaltlichen Kennwerten (H_4).

Tabelle 65 Regressionsanalyse (GESAMTBEWERTUNG)

				1 MP	
	Kennwerte	Df	Korr. R^2	ρ	F
Gesamtmodell	Struktur	4	0.55	0.04	4.41
	Semantik	3	0.10	0.31	1.41

MP = Messzeitpunkt

Die Korrelationskoeffizienten ergaben bei den strukturellen Kennwerten signifikante Zusammenhänge von r = 0.69 bei Surface und von r = 0.54 bei Structural Matching. Dies weist darauf hin, dass wenn die Schüler einen ähnlichen Expertisegrad vorwiesen, wie dem der Lehrkraft, sie eine bessere Gesamtbewertung erhielten. Der Concept Matching Wert von r=0.52 darf nicht interpretiert werden, da nicht auszuschließen ist, dass er zufällig zustande kam (vgl. Tabelle 65).

Tabelle 66 Korrelationskoeffizienten (1. Messzeitpunkt)

		Gesamtmodell
	Kennwerte	
Struktur	Surface Matching	0.69 (*)
	Graphical Matching	0.48
	Structural Matching	0.54 (*)
	Gamma Matching	0.39
Semantik	Concept Matching	0.52 (*)
	Propositional Matching	0.36
	Balanced Semantic Matching	0.18

8.4.4 Post-Hoc-Analyse

Die durchschnittliche Textlänge der Schülerantworten lag bei 539 Wörtern. Diese variierten von 288-741 Wörter (vgl. Tabelle 67).

Tabelle 67 Deskription der Wortanzahl

	AM	**SD**	**MD**	**Range**
Wortanzahl	538.75	173.19	571	288- 741

Die Post-Hoc-Analyse ergab vor der Intervention einen signifikanten Korrelationskoeffizieten von r = 0.84 und nach der Intervention einen signifikanten Korrelationskoeffizieten von r = 0.80. Dies zeigt, dass sich die Lehrkraft durch die Textlänge beeinflussen ließ. Nach der Intervention nahm dieser Effekt minimal ab. Es kann jedoch nicht ausgeschlossen werden, dass die längeren Schülertexte wirklich besser waren als die kürzeren Schülertexte.

Tabelle 68 Wortanzahl und Gesamteindruck

	1 MP	**2 MP**
Gesamtbewertung	0.84 (*)	0.80 (*)

MP = Messzeitpunkt

8.4.5 Güte der Musterlösung

Die Untersuchung der in der Musterlösung abgebildeten Kriterien ergab, dass alle Kriterien in der Musterlösung enthalten waren (vgl. Tabelle 69). Demzufolge kann auf eine hohe Validität der Musterlösung auf Grundlage der Kriterien geschlossen werden.

Tabelle 69 Güte der Musterlösung

	Kriterium	**vorhanden**	**teilweise vorhanden**	**nicht vorhanden**
Aufgabe 1	Nennung von Autor, Quelle und Schwerpunkt des Textes	X		
	Zusammenfassung der Hauptthesen	X		
	Verwendung von eigenen Worten	X		
	Kennzeichnung von Zitaten	X		
	Konjunktiv	X		
Aufgabe 2	Jüdische Erwartung	X		
	Erwartung von Jesus	X		
	Fazit/ Vergleich	X		

8.5 Ergebnisse der Untersuchungen im Fach Kunst

8.5.1 Bewertungskriterien

Die Bewertungskriterien umfassen eine Beschreibung, eine Analyse sowie eine Interpretation der im Bild gesehenen Komponenten (vgl. Tabelle 70-71). Die Kriterien die das Bild beschreiben, unterteilen sich in die einzelnen Bildbereiche (untere Hälfte sowie obere Bildhälfte und jeweils in einen rechten sowie linken Bereich) des Betrachters. Die Maßstäbe der Analyse des Bildes unterteilen sich in Farb-, Raumanalyse und Komposition. Die Kriterien der Interpretation umfassen: Eigenständigkeit, Aufbau und Originalität. Dabei ergeben sich bei der Beschreibung sowie der Interpretation jeweils acht Punkte. Die Analyse ergab insgesamt sechs Punkte.

Tabelle 70 Bewertungskriterien im Unterrichtsfach Kunst (Beschreibung)

Beschreibung	
untere Hälfte	grünlich-blaue Fläche (= See eins) grenzt ab; darunter rote länglich-kurvige Fläche (= Gewand/ Gestalt) von rechts bis fast ganz links; Hüftknick der Gestalt = Bildmitte
unten rechts	vier Gesichter/ Köpfe – drei fantastische (= Dämonen), ein bärtig-schreckverzerrtes Gesicht der Gestalt (= Mensch); rechts + links neben Kopf der Gestalt aus Gewandöffnungen Hand und Arm; [genauere Beschreibung der Dämonen...]
unten links	links neben Hüfte der Gestalt bzw. zwischen Beinen zwei Dämonen genauere Beschreibung...]; über dem oberen Bein der Gestalt ein weiterer Dämon […]; links davon weiterer Dämon in flauschigem Fell […] darunter zwei schalentierartige Dämonenköpfe ins Gewand verbissen und links davon ein Kinderkopf mit bläulichem Gegenstand im Mund und violetter Kralle über der Stirn; darüber eine sehr große Dämonenfratze (ein Drittel der Bildhöhe) zwischen deren weit stehenden Augen ein weiteres mit Zähnen bewehrtes Maul sich öffnet woraus wiederum ein Kopf mit zwei Hörnern und einem roten Auge stößt; im weit aufgerissenen regulären Maul der Fratze zwischen scharfen Zähnen auf deren Zunge lauert ein ebenfalls mit Zähnen bewehrter Dämon als Mischung aus Katze und Fisch; darüber Formen wie Gesträuch + ein echsenartiger Kopf sowie eine Spinne in der Felsfarbe auf einer felsigen Mauerform (ein Fünftel des Bildes vom linken Rand über die volle Höhe); wie hier die Mauer oben und unten im Bild verbindet, so am rechten Rand ein baumartiges Gewächs mit Getier-/spinnenartiger Krone;
obere Bildhälfte	mittig auf Achse des Hüftknicks des Gestalt teilt ein rechteckiger vor allem links stark bewachsener Felsen die obere Hälfte; zwei Kreisformen (= Augen) und eine Öffnung unten (= Mund) auf diesem lassen Gesicht erkennen; Gewächse links mit Ranken etc. lassen nackten weiblichen Körper erkennen; oben links blauer Himmel nach rechts gelblich-grüne Gewitterwolken
links oben	großer Felsen links der Mauer am linken Bildrand, rechts davon kleinerer Felsen mit baumartigem Gewächs; zwischen beiden Überschneidungen (= Bach)
rechts oben	im unteren Bereich grün-braune Fläche (= Ufer) nach rechts verlaufend; darüber ein gelb-grünliche bzw. beige Fläche (= See zwei); darüber eine bläuliche schmale/ variierende Fläche (= entfernte Berge); auf dem unteren Ufer links ein Art Säule, rechts davon erhöht eine Art Baum mit hauchdünnem Stamm und dunkler Krone, rechts davon eine Säule auf deren Kapitell ein Gekreuzigte zu stehen scheint […]
Bewertungskriterien	16 von den 26 genannten Bildgegenständen = 8 Punkte (insgesamt) *Wird der Heilige Antonius im Bild nicht entdeckt, wird ein halber Punkt abgezogen.

Tabelle 71 Bewertungskriterien im Unterrichtsfach Kunst (Analyse und Interpretation)

Analyse	
Farbanalyse:	- Farbwahl: blaue Mauer + Dämonen links, rote Gestalt und „Blatt“ am Baum vorne rechts, Dämonen violett, gelb-grünlich und bläulich vorne rechts, See 1 bläulich-türkis, Felsen und Landzunge darüber grün-bläulich bis grün-braun, See 2 beige und gelb-grünlich, Bergkette in verblasstem Blau; Wolken darüber Senf gelb mit grünlich-blau; Himmel oben rechts blau zum Horizont aufhellend bis weiß - Farbauftrag: Vordergrund deckend bis pastos […] (Decalcomanie + Grattage), beide Seen und Bergkette lasierend […]; Felsen und Landzunge mittig teils-teils […] (Decalcomanie + Frottage) - Farbkontraste: Komplementärkontrast […], Qualitätskontrast […], Quantitätskontrast […], Warm-Kalt-Kontrast […], Hell-Dunkel-Kontrast […]
Raumanalyse: Komposition:	- Größenabnahme […], Überschneidung […], Farbperspektive […], Luftperspektive - symmetrisch in zwei Hälften: oben/ unten […], rechts/links […] - Kurvenform durch Gestalt […] - Unterteilung in Fünftel durch Mauer links
Bewertungskriterien	zwölf Analysebeispiele (sechs aus Farbe, vier aus Raum, zwei aus Komposition) = 6 Punkte (insgesamt)
Interpretation	
Bewertungskriterien	Eigenständigkeit, Aufbau, Originalität... = 8 Punkte (insgesamt)

Die Lehrkraft änderte die Kriterien sowie die Musterlösung nach der Intervention nicht.

8.5.2 Ergebnisse

Die deskriptiven Analysen veranschaulichen, dass die Lehrkraft im Unterrichtsfach Kunst die Schülerergebnisse nach der Intervention durchschnittlich niedriger einschätzte. Der Vergleich der Standardabweichungen und der Ränge zeigt, dass die Punktestreuung nach der Intervention höher ist. Insgesamt sind die jeweiligen Einschätzungen zwischen den beiden Messzeitpunkten relativ stabil.

Tabelle 72 Deskription der Bewertungen

1 MP				**2 MP**			
AM	SD	MD	Range	AM	SD	MD	Range
10.03	2.38	10.50	4- 14	9.31	2.50	9.00	3- 14

MP = Messzeitpunkt

Die folgende Tabelle stellt die Ähnlichkeitswerte dar, welche die Nähe der Lernerergebnisse mit den Referenzmodellen abbildet. Die strukturellen

Ähnlichkeiten der Lernergebnisse mit den Musterlösungen sind größer als die semantischen Ähnlichkeitskennwerte. Der hohe Surface-Wert zeigt, dass die einzelnen Schülermodelle eine ähnliche Anzahl an Verbindungen in ihren Modellen hatten, wie die in den jeweiligen Referenzmodellen. Dies lässt einen ähnlichen Expertisegrad der durchschnittlichen Schülerleistungen mit den Außenkriterien vermuten.

Tabelle 73 Deskription der Ähnlichkeitskennwerte (N = 32)

	Kennwerte	**ML**		**GM**	
		AM	SD	AM	SD
Struktur	Surface Matching	0.72	0.13	0.71	0.19
	Graphical Matching	0.67	0.22	0.74	0.15
	Structural Matching	0.61	0.30	0.54	0.27
	Gamma Matching	0.66	0.15	0.65	0.14
Semantik	Concept Matching	0.10	0.07	0.40	0.11
	Propositional Matching	0.01	0.02	0.20	0.11
	Balanced Semantic Matching	0.05	0.12	0.50	0.24

ML = Musterlösung; GM = Gesamtmodell

8.5.3 Bewertungsstabilität

Die Interkoderreliabilität zwischen den einzelnen Messzeitpunkten ergab einen signifikanten Korrelationskoeffizienten (Spearman) von $r = 0.73$ (bei $N = 32$ Lernenden), was auf eine niedrige Bewertungsübereinstimmung hindeutet (vgl. Tabelle 74). Demzufolge genügte die Messzuverlässigkeit nicht den wissenschaftlichen Standards (vgl. Bühner, 2004, S.129).

Tabelle 74 Interkoderreliabilität zwischen den Messzeitpunkten

	r
Bewertung insgesamt	0.73 (*)

8.5.4 Hypothesenprüfende Darstellung

Die Regressionsanalysen ergaben keine signifikanten Zusammenhänge zwischen den durch die Lehrkraft vergebenen Leistungspunkten und den strukturellen und semantischen Ähnlichkeitsmaßen. Weder die Hypothese H_4 noch H_5 darf angenommen werden. Die Gesamtbewertung lässt sich weder durch die strukturellen noch durch die semantischen Kennwerte abbilden.

8.5.5 Post-Hoc-Analyse

Durchschnittlich umfassten die einzelnen Schülertexte 841 Wörter. Der kürzeste Schülertext umfasst 327 Wörter und der längste 1444 Wörter (vgl. Tabelle 75).

Tabelle 75 Deskription der Wortanzahl (N = 29)

	AM	SD	MD	Range
Wortanzahl	841.13	232.66	839.50	327- 1444

Die Post-Hoc-Analyse zur Überprüfung, inwiefern die Textlänge einen Einfluss auf die Gesamtbewertung hatte, ergab beim ersten Messzeitpunkt einen signifikanten Zusammenhang (nach Spearman) von r = 0.60 und beim zweiten Messzeitpunkt einen signifikanten Zusammenhang von r = 0.74. Dies lässt vermuten, dass sich die Lehrkraft bei der Bewertung der Schülerleistungen durch die Textlänge beeinflussen ließ. Dieser Effekt nahm nach der Intervention zu (vgl. Tabelle 76).

Tabelle 76 Wortanzahl und Gesamteindruck

	1 MP	2 MP
Gesamtbewertung	0.60 (*)	0.74 (*)

MP = Messzeitpunkt

8.5.6 Güte der Musterlösung

Die Überprüfung, inwiefern alle in den Kriterien enthaltenen Aspekte in der Musterlösung enthalten sind, ergab Folgendes. Bei der Beschreibung waren die Kriterien: untere Bildhälfte, unten links sowie die obere Bildhälfte vorhanden. Die Kriterien der unteren, der links-oberen sowie der rechts-oberen Bildhälfte waren teilweise vorhanden. Bei der Analyse war das Kriterium Raumanalyse vorhanden. Die Kriterien Farbanalyse und Komposition waren teilweise in der textbasierten Musterlösung enthalten (vgl. Tabelle 77). Das Kriterium Interpretation war nicht in der Musterlösung vorhanden. Da nicht alle inhaltlichen Kriterien in der Musterlösung abgebildet sind, lässt dies keinen inhaltsvaliden Vergleich als Außenkriterium zu. Insbesondere die letzte Aufgabenstellung bezüglich der Interpretation kann demzufolge nicht mit den einzelnen Schülertexten verglichen werden.

Tabelle 77 Güte der Musterlösung

		vorhanden	**teilweise vorhanden**	**nicht vorhanden**
Beschreibung	untere Hälfte	X		
	unten rechts		X	
	unten links	X		
	obere Bildhälfte	X		
	links oben		X	
	rechts oben		X	
Analyse	Farbanalyse		X	
	Raumanalyse	X		
	Komposition		X	
Interpretation				X

8.6 Ergebnisse der Interviews

Im folgenden Abschnitt werden die Ergebnisse der Interviewauswertungen mithilfe der Qualitativen Inhaltsanalyse dargestellt. Dieser unterteilt sich in die Ergebnispräsentation der induktiven sowie der deduktiven Auswertungen. Erstes fokussiert die Beantwortung der Fragestellungen, welche Bewertungskriterien Lehrkräfte in den vorliegenden Untersuchungen bei der Bewertung textbasierter Schülerleistungen konkret anwendeten sowie was deren Vorgehensweisen von der Erstellung der Klausuren bis hin zur Notengebung waren. Diese Fragestellungen wurden auf Grundlage induktiver Analysen beantwortet. Zweites fokussiert die Bestimmung der Bewertungsqualität, die in der vorliegenden Studie teilgenommenen Lehrkräfte. Die Beantwortung dieser Fragestellung wurde mithilfe deduktiver Analysen beantwortet.

Zunächst erfolgen eine Darstellung der zentralen Ergebnisse und daran eine Darstellung der Güte der (induktiven und deduktiven) Analysen mithilfe der Reliabilitätsbestimmungen. Dabei erfolgt eine nach Fächern getrennte Darstellung zunächst der induktiven Auswertungen (nach Kriterien und Vorgehensweise getrennt) sowie daran anschließend der deduktiven Auswertungen hinsichtlich der Bewertungsqualität.

8.6.1 Bewertungskriterien

Zur Beantwortung der ersten Fragestellung, welche Kriterien in den vorliegenden Untersuchungen Lehrkräfte bei der Bewertung textbasierter Schülerleistungen anlegten, ergaben sich die im Folgenden dargestellten Ergebnisse.

8.6.1.1 Bewertungskriterien im Fach Biologie

8.6.1.1.1 Erste Lehrkraft

Folgende Tabelle (78) stellt die Ergebnisse aus den Interviews der ersten Lehrkraft in Unterrichtsfach Biologie dar. Das erste Interview wurde aus technischen Gründen nicht aufgenommen. Deswegen wurden diese rekonstruiert. Die induktive Auswertung der Bewertungskriterien der ersten Lehrkraft in Biologie ergab beim ersten Messzeitpunkt eine Unterteilung in Inhalt und Form. Beide Kategorien lehnten sich an die Kompetenzbereiche Wiedergabe, Verständnis und Transfer an. Sie wurden als Bewertungskriterien der Oberkategorie zusammengefasst. Die Analyse des zweiten Messzeitpunktes ergab eine Unterteilung der Bewertungskriterien in Erwartungshorizont, Sprache und Gliederung, Materialbenutzung, Menge, Antwort, Argumentation und Begründung am Material sowie eigene Materialbenutzung. Die Lehrkraft berücksichtigte die soziale Bezugsnorm indem sie die Klassenergebnisse zur Kenntnis nahm. Der Erwartungshorizont lehnte sich an die zuvor behandelten Unterrichtsinhalte an. Die Sprache und Gliederung umfassten folgende Aspekte: eine abgeschlossene Darstellung unter Berücksichtigung der Regeln der deutschen Sprache, ein Hinführen zum Thema, das Bearbeiten der Aufgabenstellung sowie das Fazit und die Zusammenfassung. Dabei waren die Verwendung von Fachtermini sowie der Bezug zu den Arbeitsmaterialien wichtig. Schließlich umfassten diese Kriterien den kompletten Satzbau sowie die Anwendung der Rechtschreib- und Grammatikregeln. Das Kriterium der Materialbenutzung umfasste den inhaltlichen sowie schlussfolgernden Bezug zu den Materialien. Das Kriterium Menge fokussierte eine maximale Begrenzung auf insgesamt zwei Seiten. Das Kriterium Antwort umfasste eine eindeutige Begründung der eigenen Stellungnahme. Argumentation und Begründung am Material meinten das Einbringen von Pro und Contra Argumenten auf detaillierte, konkrete sowie inhaltliche richtige Weise

sowie das Einbringen von Beispielen und Zusammenhängen. Das eigene Materialbeispiel ließ im Gegensatz zu dem eben genannten eine oberflächliche Auseinandersetzung zu.

Tabelle 78 Bewertungskriterien der ersten Lehrkraft im Unterrichtsfach Biologie

MP	Zeile	Kategorie	Bemerkung	Bestimmung der Oberkategorie
1	1-3	Inhalt	In Anlehnung an die Kompetenzbereiche: Wiedergabe, Verständnis, Transfer	BK
1	1-3	Form	In Anlehnung an die Kompetenzbereiche: Wiedergabe, Verständnis, Transfer	BK
2	1-4	Erwartungshorizont	In Anlehnung an die Unterrichtsinhalte	BK
2	5-24	Sprache und Gliederung	Abgeschlossene Darstellung; Regeln der deutschen Sprache, Hinführen zum Thema; (fachlich-inhaltliches) Bearbeiten der Aufgabenstellung; Fazit, Zusammenfassung; Verwendung von Fachtermini; Bezug zu Arbeitsmaterialien; kompletter Satzbau; Anwenden der Rechtschreib- und Grammatikregeln	BK
2	25-38	Materialbenutzung	Inhaltlicher sowie schlussfolgender Bezug zu den Materialien	BK
2	39-46	Menge	Begrenzung auf maximal zwei Seiten	BK
2	47-70	Antwort	Eindeutige Begründung der eigenen Stellungnahme	BK
2	71-110	Argumentation und Begründung am Material	Einbringen (detaillierter, konkreter sowie inhaltlich richtiger) Pro und Contra Argumente; Beispiele und Zusammenhänge	BK
2	11-124	Eigenes Materialbeispiel	Einbringen eines jeweils (detaillierten, konkreten sowie inhaltlich richtigen) Pro und Contra Argumente; Oberflächlichkeit zugelassen	BK
2	90-91	Soziale Bezugsnorm	Kenntnisnahme der Ergebnisse der Klasse	BN

BK = Bewertungskriterien; BN = Bezugsnorm

8.6.1.1.2 Zweite Lehrkraft

Die induktive Auswertung der Bewertungskriterien der zweiten Lehrkraft ergab beim ersten Messzeitpunkt eine Unterteilung in Inhalt und Kompetenz. Als Oberkategorie wurden diese als kriteriale Bewertungskriterien zusammengefasst. Der zweite Messzeitpunkt zeigte dieselbe Unterteilung in inhaltliche Punkte und Kompetenz sowie deren Gewichtung (vgl. Tabelle 79).

Tabelle 79 Bewertungskriterien der zweiten Lehrkraft im Unterrichtsfach Biologie

MP	Zeile	Kategorie	Bemerkung	Bestimmung der Oberkategorie
1	1-6	Inhalt	Inhalt und deren Gewichtung	kriteriale Bewertungskriterien
1	1-6	Kompetenz	Kompetenz und deren Gewichtung	kriteriale Bewertungskriterien
2	1-13	Inhaltliche Punkte/ Fakten	Inhalt und deren Gewichtung	kriteriale Bewertungskriterien
2	8-9	Kompetenz	Kompetenzpunkte und deren Gewichtung	kriteriale Bewertungskriterien

Bei der Analyse des Nachinterviews, in dem gefragt wurde, nach welchen Aspekten genau in den Schülerleistungen gesucht wurde - um die jeweiligen Kriterien zu erfüllen - ergaben die induktiven Auswertungen eine Unterteilung in folgende Kategorien: Ecstasy, Fazit, Alkohol, LSD, selbst gewähltes Beispiel und Kompetenzpunkte. Bei dem inhaltlichen Kriterium Ecstasy achtet die Lehrkraft in den Schülerleistungen auf den inhaltlich richtigen Zusammenhang, welcher auch über den Erwartungshorizont hinausgehen konnte. Hier musste jeweils mindestens ein aus dem Erwartungshorizont aufgeführter Punkt angesprochen sein (siehe Tabelle 35). Das Fazit fokussierte die erwartete Wirkung sowie die Nebenwirkung (hinsichtlich gesundheitlicher Schäden und die Wirkung auf den Körper). Diese Auswirkungen sollten zumindest erwähnt. Bei den Kriterien Alkohol und LSD sollte mindestens eine der Begrifflichkeiten aus dem Kriterienkatalog angesprochen sein. Die Kriterien in Bezug auf das selbst gewählte Beispiel schlossen die beabsichtigte Wirkung und Nebenwirkung ein. Die Kompetenzpunkte beabsichtigten eine sinnvolle Gliederung, die Beschränkung auf das Wesentliche und den Einbezug aller Materialien, richtige Verwendung von Fachtermini und einen angemessenen Ausdruck. Die Lehrkraft gab an, dass die Bestimmung der Punkteverteilung für die Kompetenzpunkte durch die Lehrkraft subjektiv erfolgte. Dies kann Tabelle 80 entnommen werden.

Tabelle 80 Bewertungskriterien der zweiten Lehrkraft im Unterrichtsfach Biologie (Nachinterview)

Zeile	Kategorie	Bemerkung	Bestimmung der Oberkategorie
1-53; 84-87	Ecstasy	Inhaltlich richtiger Zusammenhang, auch über den Erwartungshorizont hinaus; Jeweils mindestens ein aus dem Erwartungshorizont aufgeführter Punkt muss angesprochen sein	Bewertungskriterien für Ecstasy
54-63	Fazit	Erwartete Wirkung, Nebenwirkung (gesundheitliche Schäden, Wirkung auf Körper) ansprechen	Gewichtung der Punkte beim Fazit
67-83	Alkohol	Mindestens einer der Begrifflichkeiten aus dem Kriterienkatalog (zu Alkohol) muss angesprochen werden	Bewertungskriterien für Alkohol
88-99	LSD	Mindestens einer der Begrifflichkeiten aus dem Kriterienkatalog (zu LSD) muss angesprochen werden	Bewertungskriterien für LSD
100-122	selbst gewähltes Beispiel	Beispiel sowie beabsichtige Wirkung und Nebenwirkung	Bewertungskriterien für selbst gewähltes Beispiel
123-169	Kompetenz-punkte	Sinnvolle Gliederung (Erkennbarkeit von mindestens drei (bzw. vier) Drogen; Beschränkung auf das Wesentliche sowie Einbezug aller Materialien; Richtige Verwendung von Fachtermini sowie angemessener Ausdruck; subjektive (von der Lehrkraft festgelegte) Punkteverteilung	Bewertungskriterien für Kompetenzpunkte

8.6.1.2 Bewertungskriterien im Fach Deutsch

Die induktive Auswertung der Bewertungskriterien der Lehrkraft im Unterrichtsfach Deutsch ergab beim ersten Messzeitpunkt eine Unterteilung in Inhalt und Sprache sowie deren Gewichtung. Beide Aspekte wurden als Oberkategorie Bewertungskriterien genannt. Die Analyse des zweiten Messzeitpunkts ergab dieselbe Unterteilung. Das inhaltliche Kriterium fokussierte das Benennen des Themas, das Vorhandensein von Überleitungssätzen, den Argumentationsaufbau sowie Beispielen, einem Schluss und Ausblick. Das sprachliche Kriterium umfasste die Qualität der Überleitungssätze, die stilistische Qualität der Ausdrucksweise und Formulierung und schließlich die Qualität der Rechtschreibung, Zeichensetzung und Grammatik.

Tabelle 81 Bewertungskriterien im Unterrichtsfach Deutsch

MP	Zeile	Kategorie	Bemerkung	Bestimmung der Oberkategorie
1	7-11	Inhalt	Inhalt und deren Gewichtung	Bewertungskriterien
1	7-11	Sprache	Sprache und deren Gewichtung	Bewertungskriterien
2	13-18	Inhalt	Inhalt und deren Gewichtung	Bewertungskriterien
2	13-18	Sprache	Sprache und deren Gewichtung	Bewertungskriterien
2	13-18	Form	Form und deren Gewichtung	Bewertungskriterien
2	19-33	Inhalt	Nennung des Themas	Bewertungskriterien
2	34-39	Inhalt	Vorhandensein von Überleitungssätzen	Bewertungskriterien
2	40-51	Inhalt	Argumentationsaufbau	Bewertungskriterien
2	52-59	Inhalt	Beispiele	Bewertungskriterien
2	60-71	Inhalt	Enthält Schluss einen Ausblick	Bewertungskriterien
2	72-81	Sprache	Qualität der Überleitungssätze	Bewertungskriterien
2	82-97	Sprache	Stilistische Qualität (Ausdrucksweise und Formulierung)	Bewertungskriterien
2	98-106	Sprache	Qualität der Rechtschreibung, Zeichensetzung und Grammatik	Bewertungskriterien

8.6.1.3 Bewertungskriterien im Fach Religion

8.6.1.3.1 Erste Lehrkraft

Die induktive Auswertung der Bewertungskriterien der ersten Lehrkraft in Religion ergab beim ersten Messzeitpunkt eine Unterteilung in Organisation des Textes sowie in Unterrichtsinhalt als Vergleichsmaßstab. Beide Aspekte beeinflussen die Formulierung des Erwartungshorizontes. Zweites wirkt sich zudem auf die Punkteverteilung aus. Das erste Kriterium fokussiert den ersten Aufgabenbereich und das zweite Kriterium den zweiten Aufgabenbereich.

Die induktive Auswertung des zweiten Messzeitpunktes ergab eine Unterteilung in: Inhalt, Aufbau, Ausdrucksweise, Textzusammenfassung sowie Zusammenfassung und Unterteilung in zweite und dritte Aufgabenstellung. Das Inhaltskriterium schließt den thematischen Bezug sowie die inhaltliche Richtigkeit ein. Das Kriterium des Aufbaus umfasste die Argumentationsweise. Die Ausdrucksweise beinhaltet Satzbau, Formulierung, Einbezug von Fachbegriffen, Verständlichkeit sowie Plausibilität. Die Textzusammenfassung beinhaltet die Plausibilität sowie die inhaltliche Richtigkeit. Die Zusammenfassung schließt den Inhalt und die Argumentationsstruktur ein. Die zweite Aufgabenstellung fokussierte den inhaltlichen Bezug zur Aufgabenstellung und der dritte Aufgabenbereich umfasste

die individuelle, inhaltliche Ausführung. Alle Kategorien wurden als Oberkategorie Bewertungskriterien bestimmt.

Tabelle 82 Bewertungskriterien der ersten Lehrkraft im Unterrichtsfach Religion

MP	Zeile	Kategorie	Bemerkung	Bestimmung der Oberkategorie
1	1-12	Organisation des Textes	Formulierung und Orientierung am Erwartungshorizont	Aufgabenbereich 1
1	9-14	Unterrichtsinhalt als Vergleichsmaßstab	Formulierung des Erwartungshorizontes sowie Punkteverteilung	Aufgabenbereich 2
2	5-15; 54-76	Inhalt	Thematischer Bezug sowie inhaltliche Richtigkeit	Bewertungskriterien
2	5-15	Aufbau	Argumentationsweise	Bewertungskriterien
2	22-28; 41-60	Ausdrucksweise	Satzbau, Formulierung, Einbezug von Fachbegriffen, Verständlichkeit, Plausibilität	Bewertungskriterien
2	29-36	Textzusammenfassung	Plausibilität, inhaltliche Richtigkeit des Inhaltes	Bewertungskriterien
2	78-94	Zusammenfassung	Inhalt und Argumentationsstruktur	Bewertungskriterien
2	96-100	2. Aufgabenstellung	Inhaltlicher Bezug zur Aufgabenstellung	Bewertungskriterien
2	102-115	3. Aufgabenbereich	Individuelle inhaltliche Ausführung	Bewertungskriterien

8.6.1.3.2 Zweite Lehrkraft

Die induktive Auswertung der Bewertungskriterien der zweiten Lehrkraft in Religion ergab beim ersten Messzeitpunkt eine Unterteilung in Sprache und Inhalt. Die Sprache umfasste die Qualität des Ausformulierens. Der Inhalt lehnte sich an den Erwartungshorizont an. Für beide Aspekte wurden als Oberkategorie kriteriale Bewertungskriterien gewählt.

Die induktive Analyse des zweiten Messzeitpunkts beinhaltete die Bewertungsgrundlage, die generelle Bewertung, die sprachliche Leistung, die Grammatik, den Inhalt, die Punkteverteilung sowie das Fazit. Dabei orientierte sich die Bewertungsgrundlage am Erwartungshorizont, der Musterlösung und den behandelten Unterrichtsinhalten. Die generelle Bewertung fokussierte Ausdruck sowie die sprachliche Leistung während sich das Kriterium der sprachlichen Leistung auf die Umgangssprache, Fachbegriffe, Rechtschreibung sowie Zeichensetzung bezog. Das grammatikalische Kriterium umfasste die Anzahl der Fehler geteilt durch die Anzahl der Wörter im Text insgesamt. Beim

Inhaltskriterium war z. B. die Nennung von Autor, Quelle und Schwerpunkt des Textes. In den Zeilen 57-107 erklärte die Lehrkraft, wie sie die Punkte am konkreten Beispiel verteilt. Dabei stellte sich heraus, dass die Schüler die Transferaufgabe (Aufgabenbereich 3) nicht lösen konnten, weil diese nach den Vermutungen der Lehrkraft zu kompliziert gedacht hatten. Deswegen wurde diese Aufgabe von dieser Lehrkraft nicht bewertet (vgl. Tabelle 83, Zeile: 57-107). Das Fazit umfasste eine kurze Zusammenfassung sowie den Bezug zur Argumentationsweise, wobei keine neuen Aspekte hineinfließen sollten. Als Oberkategorien wurden Bewertungskriterien sowie die Anwendung der Bewertungskriterien bestimmt.

Tabelle 83 Bewertungskriterien der zweiten Lehrkraft im Unterrichtsfach Religion

MP	Zeile	Kategorie	Bemerkung	Bestimmung der Oberkategorie
1	1-14	Sprache	Qualität des Ausformulierens (Formulierung, Umgangssprache)	Kriteriale Bewertungskriterien
1	1-14	Inhalt	In Anlehnung an den Erwartungshorizont	Kriteriale Bewertungskriterien
2	4-11	Bewertungsgrundlage	Erwartungshorizont, Muster-lösung, Unterrichtsinhalte	Bewertungskriterien
2	12-19	Generelle Bewertung	Ausdruck, sprachliche Leistung	Bewertungskriterien
2	20-38	Sprachliche Leistung	Umgangssprache, Fachbegriffe, Rechtschreibung, Zeichensetzung	Bewertungskriterien
2	39-48	Grammatik	Anzahl der Fehler/Anzahl der Wörter	Bewertungskriterien
2	49-51	Inhalt	z. B. Nennung von Autor, Quelle und Schwerpunkt des Textes	Bewertungskriterien
2	57-107	Punkteverteilung	Erklärung am konkreten Beispiel	Anwendung der Bewertungskriterien
2	108-116	Fazit	Kurze Zusammenfassung sowie Bezug zur Argumentationsweise keine neuen Aspekte	Bewertungskriterien

8.6.1.4 Bewertungskriterien im Fach Kunst

Die induktive Auswertung des ersten Messzeitpunktes ergab eine Unterteilung der Bewertungskriterien in die Aufgabenbereiche 1, 2 und 3. Diese orientierten sich am Stuttgarter Modell, welches die Vorgehensweise sowie die Trennung der einzelnen Aufgabenbereiche (Beschreiben, Analysieren sowie Interpretieren) festlegt. Der erste Aufgabenbereich umfasste die Bildbeschreibung, der zweite Aufgabenbereich

fokussierte die Bildanalyse und der dritte Aufgabenbereich die Interpretation. Hier wurden als Oberkategorien Bewertungskriterien sowie Klausurvorgabe festgelegt. Die induktiven Analysen des zweiten Messzeitpunktes ergaben bei den Oberkategorien eine Unterteilung in Bewertungskriterien, Beschreiben, Analyse sowie Interpretation. Beim Beschreiben folgte eine zusätzliche Oberkategorie Bildeinteilung für die Bildbeschreibung. Zudem wurden die Oberkategorien Bewertungskriterien und Aspekte außerhalb des Kriterienkatalogs sowie Musterlösung bestimmt. Für die Bewertungskriterien erfolgte eine kategoriale Einteilung in Inhalte der Aufgabenstellung welche, das Beschreiben, das Analysieren sowie das Interpretieren umfassten, und das Beschreiben, welches die Nennung und Differenzierung fokussierte, worauf die Analyse und Interpretation aufbaute. Bei der Oberkategorien Beschreiben erfolgte eine kategoriale Einteilung in Umfang, schlüssiger Aufbau, obere Bildhälfte sowie die Bildhälfte rechts oben. Der schlüssige Aufbau meinte, dass die Interpretation auf die Bildbeschreibung aufbaute. Die Kriterien der Bildanalyse umfassten folgende Kategorien: Umsetzung der Textanalyse, Farbwahl, und -kontraste, Raumanalyse, welche sich unterteilte in Größenabnahme, Luftperspektive, Überschneiden, sowie schließlich die Komposition. Die Umsetzung der Textanalyse fokussierte die Analyse sowie das Anführen von Beispielen. Die Farbwahl umfasste die Nennung der Farben nach der Bezugsnorm des Malers sowie die Nennung der im Bild enthaltenen Primär- und Sekundärfarben. Außerdem meinte es die Bezugnahme zu den Farbvariationen im Bild, um damit für die Interpretation die Grundlage zu legen. Die Interpretation umfasste die Eigenständigkeit, Originalität sowie Aufbau und Punkteverteilung. Die Eigenständigkeit wurde in der Ehrlichkeit und dem Mut der Schüler, ihre Gedanken und Ideen zu formulieren gemessen. Hier gab es Abzüge, wenn diese nicht in den Aufbau und die Struktur einbezogen wurden. Die Originalität fokussierte den Einbezug neuer Gedanken, sowie das „Sprünge gehen“. Mit Aufbau war die Nachvollziehbarkeit des Themas gemeint. Die Oberkategorie der Bildeinteilung hinsichtlich der Bildbeschreibung wurde als Kategorie: Menge und Bildeinteilung bestimmt. Die Lehrkraft verwies darauf, dass die im Kriterienkatalog enthaltenen Schlagwörter von den Schülern in den richtigen Kontext gebracht werden mussten (vgl. Tabelle 84, Zeile: 129 - 131). Schließlich gab die Lehrkraft an, Punkte zu vergeben, wenn die Schüler mehr Aspekte fanden, als im Kriterienkatalog enthalten waren (vgl. Tabelle 85, Zeile: 147 - 150).

Tabelle 84 Bewertungskriterien im Unterrichtsfach Kunst

MP	Zeile	Kategorie	Bemerkung	Bestimmung der Oberkategorie
1	1-6	Aufgabenbereich 1	Bildbeschreibung	Bewertungskriterien
1	6-10	Aufgabenbereich 2	Bildanalyse	Bewertungskriterien
1	10-32	„Stuttgarter Modell“	Festgelegte Vorgehensweise; Trennung der einzelnen Aufgabenbereiche (Beschreiben, Analysieren und Interpretieren)	Klausurvorgabe
1	31-32	Aufgabenbereich 3	Interpretation	Bewertungskriterien
2	1-11	Inhalte der Aufgabenstellung	Beschreiben, Analysieren, Interpretieren	Bewertungskriterien
2	7-60	Beschreiben	Nennung und Differenzierung damit darauf die Analyse und Interpretation aufbauen kann	Bewertungskriterien
2	61-72	Umfang (der von den Schülern angesprochenen Aspekten aus dem Kriterienkatalog)	16 von insgesamt 26 Kriterien müssen genannt sein	Beschreiben
2	72-74	Schlüssiger Aufbau	Interpretation baut auf der Bildbeschreibung auf	Beschreiben
2	74-103	Menge und Bildeinteilung (der aus dem Kriterienkatalog abgebildeten Inhaltsaspekte im Bild)	Bildeinteilung: rechts, links, mitten unten oder anhand von Kompositionsformen, Konstanten, Linien und Schrägen	Bildeinteilung hinsichtlich der Bildbeschreibung
2	105-125	Obere Bildhälfte	Obere Bildhälfte und deren Inhalte	Beschreiben
2	129-131	Schlagwortsuche	Schlagwörter aus dem Kriterienkatalog müssen in den richtigen Kontext gebracht werden.	Vorgehensweise bei der Bewertung mith des Kriterienkatalogs
2	131-143	Umsetzung der Analyse im Text	Analyse sowie Anführung von Beispielen	Analyse
2	143-146	Anfertigen und Gegenstände in der Musterlösung	Lehrkraft benötigte mehr Zeit für das Anfertigen der Musterlösung, da auf alle Gegenstände Bezug genommen wurde.	Musterlösung

Tabelle 85 Bewertungskriterien im Unterrichtsfach Kunst

MP	Zeile	Kategorie	Bemerkung	Bestimmung der Oberkategorie
2	147-150	Ansprechen von Aspekten, die nicht im Kriterienkatalog enthalten waren	Punkte werden vergeben, wenn Schüler mehr Aspekte nennt oder findet - als im Kriterienkatalog enthalten	Aspekte außerhalb des Kriterienkatalogs
2	151-168	Rechts oben im Bild	Trennung von Interpretation und Beschreibung; Verwendung des Konjunktivs	Beschreiben
2	169-186	Farbwahl	Nennung der Farben nach der Bezugsnorm des Malers, Nennung der im Bild, enthaltenen Primär- und Sekundärfarben. Bezugnahme zu den Farbvariationen im Bild um damit für die Interpretation die Grundlage zu legen	Analyse
2	187-203	Farbkontraste	Nennung der Komplementärkontraste, die für die Interpretation gebraucht werden.	Analyse
2	204-225	Raumanalyse: Größenabnahme	Deutung der Gegenstände im Bild nach ihrer Größe sowie Einbezug der zuvor gefundenen Farben	Analyse
2	226-240	Raumanalyse: Luftperspektive	Verblassen der Farben im Hintergrund des Bildes als Hinweis der Entfernung	Analyse
2	241-259	Raumanalyse: Überschneiden	Erkennen der Gegenstände, die sich im Bild überschneiden sowie Interpretation	Analyse
2	260-272	Komposition	Erkennen der oberen und unteren Bildhälfte; kurvenförmige horizontale Linie im Bild; Unterteilung in Fünftel durch die Mauer; nicht im Kriterienkatalog enthaltener Aspekt zu Komposition: mittiger Schnitt durch den Felsen	Analyse
2	273-289	Eigenständigkeit	Ehrlichkeit und Mut der Schüler ihre Gedanken und Ideen zu formulieren; ohne Einbezug von Aufbau und Struktur gab es bei der Interpretation Abzüge	Interpretation
2	289-292	Originalität	Einbezug neuer Gedanken sowie Sprünge gehen	Interpretation
2	292-299	Eigenständigkeit	Einschränkung der Eigenständigkeit wenn in der beschreibenden Analyse bestimmte Sachen nicht gesehen oder zugeordnet wurden.	Interpretation
2	300-316	Aufbau	Nachvollziehbarkeit	Interpretation
2	317-345	Punkteverteilung	Eigenständigkeit, Aufbau, Originalität und deren individuell gehandhabten Gewichtungen	Interpretation

8.6.2 Vorgehensweisen

Zur Beantwortung der zweiten Fragestellung, wie die Lehrkräfte in den vorliegenden Untersuchungen von der Erstellung der Klausurfragen bis hin zur Notengebung vorgegangen sind, ergaben sich folgende Befunde.

8.6.2.1 Vorgehensweise im Fach Biologie

8.6.2.1.1 Erste Lehrkraft

Die induktive Analyse der ersten Lehrkraft im Unterrichtsfach Biologie ergab beim ersten Messzeitpunkt eine Orientierung sowohl am Lehrplan als auch an den Unterrichtsinhalten. Diese beiden Aspekte wurden als Oberkategorie Aufgabenentwicklung zusammengefasst.

Die induktiven Analysen des zweiten Messzeitpunktes ergaben ebenfalls eine Orientierung am Lehrplan und am Unterrichtsinhalt mit der Oberkategorie Aufgabenentwicklung. Zudem wurden folgende Oberkategorien festgelegt: Aufgabenüberprüfung, Abgabe des Erwartungshorizontes an die Schule, Überprüfung der Schülerleistungen mithilfe des Kriterienkatalogs sowie die Benotung. Die Aufgabenentwicklung betraf Inhalte sowie die Lösbarkeit und Zeit bezüglich der Aufgabenstellung. Die Abgabe des Erwartungshorizontes an die Schule erfolgte drei Tage vor Klausurtermin. Die Überprüfung der Schülerleistungen orientierte sich an den zuvor festgelegten Kriterien geleiteten Erwartungshorizont. Die Bestimmung der Note erfolgte durch den allgemeingültigen transparenten Notenspiegel an der Schule.

Tabelle 86 Vorgehensweise der ersten Lehrkraft im Unterrichtsfach Biologie

MP	Zeile	Kategorie	Bemerkung	Bestimmung der Oberkategorie
1	6-7	Lehrplan	Orientierung am LP	Aufgabenentwicklung
1	6-7	Unterrichtsinhalt	Orientierung am UI	Aufgabenentwicklung
2	125-128	Lehrplan, Unterrichtsinhalt	Orientierung am LP und am Unterrichtsinhalt	Aufgabenentwicklung
2	128-132	Aufgabenstellung	Aufgabenstellung und deren Inhalte und Anteile	Aufgabenentwicklung
2	128-134	Aufgabenstellung	Lösbarkeit und Zeit	Aufgabenüberprüfung
2	134-135	Erwartungshorizont	Festlegung und Abgabe des Erwartungshorizontes an der Schule drei Tage vor Klausurtermin	Abgabe an die Schule
2	137-140	Kriterien geleiteter Erwartungshorizont	Berücksichtigung der zuvor festgelegten Kriterien	Überprüfung der Schülerleistungen mithilfe des Kriterienkatalogs
2	141-148	Notenspiegel	Transparenz und Allgemeingültigkeit an der Schule	Benotung

8.6.2.1.2 Zweite Lehrkraft

Die induktive Analyse der zweiten Lehrkraft im Unterrichtsfach Biologie ergab beim ersten Messzeitpunkt die Oberkategorien: Klausuraufgabe sowie Bewertung. Die Klausuraufgaben schlossen die Aufgabenentwicklung und die Aufgabenstellung ein. Die Entwicklung umfasste den Prozess sowie die Vorgehensweise der Aufgabenentwicklung und deren Inhalte. Die Aufgabenstellung fokussierte das Wiedergeben und das Anwenden. Die Bewertung erfolgte durch die Punktevergabe, welche sich am Kriterienkatalog orientierte. Die induktive Auswertung des zweiten Messzeitpunkts ergab als Oberkategorien die Bewertung, die Punktevergabe, die Bewertung sowie die Klausuraufgaben. Die Punktevergabe schloss den Bewertungsprozess ein, wobei die Klausuren von der besten bis zur schlechtesten Schülerleistung sortiert wurden und daraufhin eine erneute Überprüfung der vergebenen Punkte erfolgte. Die Bewertung erfolgte auf Grundlage der Punkteverteilung und nach vorgegebenem Notenspiegel.
Die Klausuraufgaben umfassten die Aufgabenentwicklung, welche inhaltlich die Nennung, die Anwendung sowie den Transfer beinhaltete.

Tabelle 87 Vorgehensweise der zweiten Lehrkraft im Unterrichtsfach Biologie

MP	Zeile	Kategorie	Bemerkung	Bestimmung der Oberkategorie
1	7-20	Aufgabenentwicklung	Prozess sowie Vorgehensweise der Aufgabenentwicklung/ Inhalt der Aufgaben	Klausuraufgaben
1	21-29	Aufgabenstellung	Stufenbereiche (Wiedergabe und Anwendung) bei Aufgabenstellung	Klausuraufgaben
1	30-40	Punktvergabe	Bewertung nach Kriterienkatalog	Bewertung
2	14-21	Bewertungsprozess	Sortierung der Klausuren vom Besten zum Schlechtesten und erneute Überprüfung der vergebenen Punkte	Punktvergabe
2	22-26	Punktverteilung	Bewertung nach vorgegebenem Notenspiegel	Bewertung
2	27-37	Aufgabenentwicklung	Inhalt (Nennung, Anwendung, Transfer) der Aufgabenstellung	Klausuraufgaben

8.6.2.2 Vorgehensweise im Fach Deutsch

Die induktive Analyse der Lehrkraft im Unterrichtsfach Deutsch ergab beim ersten Messzeitpunkt folgende Oberkategorien: Aufgabenentwicklung, Leistungserbringung sowie Bewertung. Die Aufgabenentwicklung orientierte sich an den im Unterricht behandelten Inhalten der sozialen Netzwerke und den Elementen einer Erörterung. Die Leistungserbringung meinte die praktische Umsetzung des Aufsatzschreibens. Die Bewertung schloss den sprachlich- und inhaltlich orientierten Kriterienkatalog sowie deren Gewichtungen ein.

Beim zweiten Messzeitpunkt ergaben sich folgende Oberkategorien: Aufgabenentwicklung, Klausurvorbereitung sowie Notengebung. Die Aufgabenentwicklung bezog das Ableiten der zentralen Fragestellung aus den Unterrichtsinhalten ein. Die Klausurvorbereitung umfasste die Auswertungstransparenz. Die Kriterien waren im Unterricht entwickelt und transparent gemacht worden. Die Notengebung erfolgte nach der Intervention auf Grundlage der Punktegewichtungen. Die einzelnen Kriterien waren dabei-fünfstufig ausgeprägt.

Tabelle 88 Vorgehensweise im Unterrichtsfach Deutsch

MP	Zeile	Kategorie	Bemerkung	Bestimmung der Oberkategorie
1	16-49	Bezug zum Unterrichtsinhalt	behandelter Unterrichtsinhalt (soziale Netzwerke), Elemente einer Erörterung	Aufgabenentwicklung
1	51	Aufsatz schreiben	praktische Umsetzung	Leistungserbringung
1	56-79	Kriterienkatalog	Sprache, Grammatik, Rechtschreibung, Inhalt, Form und deren Gewichtung	Bewertung
2	121 - 126	Bezug zum Unterrichtsinhalt	Ableitung der zentralen Fragestellung aus den Unterrichtsinhalten	Aufgabenentwicklung
2	126 - 134	Auswertungs-transparenz	Entwicklung und Transparenzmachen der Bewertungskriterien im Unterricht	Klausurvorbereitung
2	135 - 158	Punktegewichtung	fünfstufiger Ausprägungsgrad der Kriterien	Notengebung

8.6.2.3 Vorgehensweise im Fach Religion

Die induktive Analyse der ersten Lehrkraft im Unterrichtsfach Religion ergab beim ersten Messzeitpunkt folgende Oberkategorien: Vorbereitung sowie Aufgabenbereich 1, 2 und 3. Die Vorbereitung umfasste die Aufgabenformulierung, welche das Heranziehen einer bereits im Unterricht behandelten Bibelstelle beinhaltete. Der erste Aufgabenbereich fokussierte die Wiedergabe, welche durch eine Verständnisfrage hervorgerufen wurde. Der zweite Aufgabenbereich umfasste das Vergleichen, mit Bezug auf das Gelernte. Der dritte Aufgabenbereich umfasste die Erörterung, welche Kreativität, Formulierung der eigenen Gedanken sowie Bezugnahme zu Aufgabenbereich eins und zwei fokussierte.

Die induktive Auswertung des zweiten Messzeitpunkts ergab folgende Oberkategorien: Aufgabenbereich 1, 2 und 3 sowie die Bewertung und der Notenspiegel. Dabei umfassten die Aufgabenbereiche die Aufgabenentwicklung, welche die Reproduktion, die Anwendung sowie den Transfer fokussierte. Die Bewertung erfolgte durch die Punktevergabe, die sich am Erwartungshorizont orientierte.

Tabelle 89 Vorgehensweise der ersten Lehrkraft im Unterrichtsfach Religion

MP	Zeile	Kategorie	Bemerkung	Bestimmung der Oberkategorie
1	21-32	Aufgabenformulierung	Heranziehen einer bereits im Unterricht behandelten Bibelstelle	Vorbereitung
1	32-34	Wiedergabe	Verständnisabfrage	Aufgabenbereich 1
1	34-38	Vergleiche	Bezugnahme zu Gelerntem	Aufgabenbereich 2
1	38-54	Erörterung	Kreativität, Formulierung der eigenen Gedanken sowie Bezugnahme zu Aufgabenbereich 1 und 2	Aufgabenbereich 3
2	117-139	Aufgabenentwicklung	Reproduktion, Anwendung, Transfer	Aufgabenbereich 1, 2 und 3
2	141-162	Punktevergabe	Punkteverteilung in Orientierung an den Erwartungshorizont	Bewertung und Notenspiegel

Die induktive Analyse der zweiten Lehrkraft im Unterrichtsfach Religion ergab beim ersten Messzeitpunkt folgende Oberkategorien: die Themenfindung, die Aufgabenentwicklung, der Erwartungshorizont und die Bewertung. Die Themenfindung umfasste den Inhalt, der sich an den Unterricht des letzten Halbjahres anlehnte. Die Aufgabenentwicklung fokussierte die Anforderungsbereiche, welche das Verstehens und Erläutern bei der Auswahl der Texte berücksichtigt. Der Erwartungshorizont schloss das Festlegen der Kriterien ein, welches durch Absprache mit Kollegen erfolgte. Die Bewertung wurde durch folgende Kategorien bestimmt: die Punkteverteilung und Notengebung, die Baucheinschätzung, die Überprüfung, die Notenpunktevergabe, der Vergleich sowie die Adaption des Erwartungshorizontes. Die Punkteverteilung und Notenvergabe schloss die Festlegung des Notenspiegels mit ein. Die Baucheinschätzung half der Gewinnung eines ersten Eindrucks. Die Überprüfung lehnte sich an den Erwartungshorizont an. Die Notenpunktevergabe umfasste eine Notenskala von 0-15 Punkten. Der Vergleich erfolgte mithilfe der Schülerleistungen untereinander sowie mit dem Erwartungshorizont. Der Erwartungshorizont wurde, sofern notwendig, im Nachhinein an die Klasse angepasst.

Die induktive Auswertung des zweiten Messzeitpunkts ergab folgende Oberkategorien: das zukünftige Konzept, die Aufgabenentwicklung, die Durchführung sowie die Bewertung. Das zukünftige Konzept meinte, dass die Klausur zukünftig vor Unterrichtsplanung sowie dessen Durchführung konzipiert

wird. Für die Aufgabenentwicklung wurden folgende Kategorien bestimmt: der Unterrichtsbezug, die Anforderungsbereiche und der Text. Der Unterrichtsbezug umfasste Gruppenpräsentationen und Tafelbilder. Die Anforderungsbereiche fokussierten das Zusammenfassen, das Verstehen sowie das Erläutern und den Transfer. Der Text orientierte sich an einer angemessenen Sprache sowie Textlänge. Die Durchführung der Klausur umfasste die Zeit, die Vorbereitung sowie das Nachfragen. Für die Bewertung wurden folgende Kategorien bestimmt: die Punkteverteilung sowie die Korrektur. Die Punkteverteilung orientierte sich an Sprache und Inhalt sowie der Gewichtung der einzelnen Aufgabenstellungen. Zudem wurden hier Bonuspunkte berücksichtigt.

Tabelle 90 Vorgehensweise der zweiten Lehrkraft im Unterrichtsfach Religion

MP	Zeile	Kategorie	Bemerkung	Bestimmung der Oberkategorie
1	34 - 39	Inhalt	in Anlehnung an den Unterricht des letzten Halbjahres	Themenfindung
1	40 - 63	Anforderungs-bereiche	Berücksichtigung der Anforderungs-bereiche (Verstehen, Erläutern) sowie Auswahl des Textes	Aufgabenent-wicklung
1	76 - 87	Festlegung der Kriterien	Absprache mit Kollegen	Erwartungs-horizont
1	87 - 89	Punkteverteilung & Notenvergabe	Festlegung des Notenspiegels	Bewertung
1	89 – 95; 105 – 110; 120; 125 – 128; 141 – 145	Baucheinschätzung	Verschaffen eines ersten Eindrucks	Bewertung
1	96; 117 - 118; 128 - 131; 145 - 146;	Überprüfung	in Anlehnung an den Erwartungshorizont	Bewertung
1	98 - 102	Notenpunkte-vergabe	Notenskala von 0-15	Bewertung
1	99 - 102; 120 - 124	Vergleich	Vergleich der Schülerleistungen unter-einander sowie mit Erwartungshorizont	Bewertung
1	133 - 150	Adaption des Er-wartungshorizontes	Anpassung an die Klasse im Nachhinein falls notwendig	Bewertung
2	122 - 123	Konzeption der Klausur	Klausurerstellung vor Unterrichtsplanung sowie dessen – Durchführung	Zukünftiges Konzept
2	124 - 126	Unterrichtsbezug	Gruppenpräsentationen, Tafelbilder	Aufgabenent-wicklung
2	127 - 134	Anforderungs-bereiche	Zusammenfassen, Verstehen, Erläutern & Transfer	Aufgabenent-wicklung
2	134 - 138	Text	angemessene Sprache & Textlänge	Aufgabenent-wicklung
2	138 - 146	Klausurdurch-führung	Zeit, Vorbereitung, Nachfrage	Durchführung
2	147 - 157	Punkteverteilung	Sprache & Inhalt und deren Gewichtung,	Bewertung
2	158 - 161	Korrektur	Aufgabe für Aufgabe, Pausen, Entspannung, Stimmung	Bewertung
2	164 - 170	Punkteverteilung	Berücksichtigung von Bonuspunkten	Bewertung
2	171 - 181	Punkteverteilung	Gewichtung der einzelnen Aufgabenstellungen	Bewertung

8.6.2.4 Vorgehensweise im Fach Kunst

Die induktiven Analysen bei der Vorgehensweise im Unterrichtsfach Kunst ergaben beim ersten Messzeitpunkt eine Unterteilung in die Oberkategorien Mustervorgehensweise, Erwartungshorizont, Punkteverteilung sowie Notengebung. Dabei unterteilte sich die Mustervorgehensweise in die Beschreibung, die Analyse sowie die Interpretation. Bei der Beschreibung gab die Lehrkraft an, dass sich die Mustervorgehensweise hinsichtlich der Bildbetrachtung insbesondere für schwächere Schüler eignete. Die Analyse umfasste eine Trennung von Beschreibung und Analyse sowie die Berücksichtigung bekannter Muster und Techniken. Die Interpretation in der Mustervorgehensweise schloss den biografischen Hintergrund sowie sieben bis acht mögliche Interpretationsoptionen ein. Der Erwartungshorizont umfasste die Bildbetrachtung wobei 16 aus 18 Details oder Gegenstände im Bild zu nennen waren. Bei der Punkteverteilung wurde eine Kategorie in Aufgabenbereich 1 und 2 bestimmt. Diese beinhalteten das Beschreiben und Analysieren sowie deren Punkteverteilungen. Für den Erwartungshorizont wurde eine Kategorie Interpretation bestimmt. Hier gab die Lehrkraft an, dass es unmöglich sei, einen idealen Erwartungshorizont zu formulieren und bei dieser Aufgabenstellung somit eine individuelle Bewertung erfolge. Bei der der Notengebung mussten mindestens fünf Punkte gegeben sein, um einen Notenpunkt zu erhalten. Die Bestimmung der Notengebung erfolgte auf Grundlage der vergebenen Punkte.

Tabelle 91 Vorgehensweise im Unterrichtsfach Kunst (erster Messzeitpunkt)

MP	Zeile	Kategorie	Bemerkung	Bestimmung der Oberkategorie
1	33-53	Beschreibung	mögliche Beschreibungsvorgehensweisen in der Bildbetrachtung; Mustervorgehensweise besonders geeignet für schwächere Schüler;	Mustervorgehensweise
1	47-53	Analyse	Trennung von Beschreibung und Analyse	Mustervorgehensweise
1	54-62	Analyse	Berücksichtigung bekannter Muster sowie der Techniken	Mustervorgehensweise
1	62-70	Interpretation	Berücksichtigung des biografischen Hintergrundes; sieben bis acht mögliche Interpretationsmöglichkeiten	Mustervorgehensweise
1	71-95	Bildbetrachtung	16 aus insgesamt 18 Details oder Gegenstände im Bild sind zu nennen; Bildkomposition; Bildbestandteile (Zuordnen wo Details oder Gegenstände im Bild gefunden wurden. Z. B. Farbkontraste)	Erwartungshorizont
1	96-104	Aufgabenbereich 1 und 2	Verteilung der Punkte hinsichtlich Aufgabenbereich 1 und 2 (Beschreibung + Analyse)	Punkteverteilung
1	98-108	Interpretation	Unmöglichkeit der Lehrkraft den idealen Erwartungshorizont zu formulieren und somit individuelle Bewertung	Erwartungshorizontes
1	109-118	Punkteverteilung	Bestimmung, ab wann es welche Note gibt; (mind. fünf Punkte müssen für einen Notenpunkt erreicht werden)	Notengebung

Beim zweiten Messzeitpunkt ergaben die induktiven Analysen eine Einteilung der Oberkategorien in Klausurentwicklung, Kriterienkatalog und Bewertung. Bei der Klausurentwicklung wurde eine Kategorie als Aufgabenentwicklung bestimmt. Dabei erfolgte eine Orientierung der einzelnen Aufgabenentwicklung am Stuttgarter Modell. Bei der Oberkategorie des Kriterienkataloges sowie der Bewertung wurde eine Kategorie als Aufwand bestimmt. In Anbetracht des Aufwandes gab die Lehrkraft an, dass der eigentliche Aufwand die Bewertung der individuellen Schülerleistungen sei, die sich für die Interpretationsmöglichkeiten im Vorfeld nicht festlegen ließen. Die Bewertung umfasste eine Kategorie der Punkteverteilung. Für die Punkteverteilung erfolgte ein Schwerpunkt auf die

Interpretation sowie die Beschreibung. Dies entspricht den Aufgabenbereichen 2 und 3.

Tabelle 92 Vorgehensweise im Unterrichtsfach Kunst (zweiter Messzeitpunkt)

MP	Zeile	Kategorie	Bemerkung	Bestimmung der Oberkategorie
2	346-366	Aufgaben-entwicklung	Orientierung am Stuttgarter Modell; Verwendung von vorhandenen Sekundärmaterialien, Verwendung von den Bildungs-standards vorgegebenen Aufgabenstellungen, der Vorgehensweise in der Klausur sowie des Erwartungshorizontes;	Klausurentwicklung
2	366-369	Aufwand	Aufwand in der Bestimmung der Interpretationsmöglichkeiten im Vorfeld sowie der eigentlichen Bewertung	Kriterienkatalog und Bewertung
2	370-381	Punktever-teilung	Schwerpunkt auf Interpretation und Beschreibung	Bewertung

8.6.3 Bewertungsqualität der Lehrkräfte

Zur Beantwortung der Fragestellung, welche Bewertungsqualität sich bei den an diesen Untersuchungen teilgenommenen Lehrkräften vorfand, ergab sich Folgendes. Die Betrachtung der Mediane ergab bei der ersten Lehrkraft im Unterrichtsfach Biologie von beiden Kodierern sowie zu beiden Messzeitpunkten eine hohe Bewertungsqualität. Die Lehrkraft in Kunst wies zu allen Messzeitpunkten (außer dem ersten Messzeitpunkt durch den ersten Kodierer) eine hohe Bewertungsqualität auf. Die Lehrkraft in Deutsch wies durch den ersten Kodierer beim ersten Messzeitpunkt eine hohe Bewertungsqualität auf. Alle anderen Lehrkräfte zeigten eine mittlere Ausprägung hinsichtlich der Bewertungsqualität, sowohl vor als auch nach der Intervention.

Tabelle 93 Bewertungsqualität aus den gemeinsamen Kategorien (Mediane)

		1 MP		2 MP	
		1. Bewerter	2. Bewerter	1. Bewerter	2. Bewerter
Biologie	1. Lehrkraft	2	2	2	2
	2. Lehrkraft	1	1	1	1
Deutsch		2	1	1	1
Religion	1. Lehrkraft	1	1	1	1
	2. Lehrkraft	1	1	1	1
Kunst		1	2	2	2

Die Beantwortung der Frage, welche Bezugsnorm Lehrkräfte bei der Analyse von Schülerleistungen anwenden, ergab folgendes. Dies war Bestandteil der Fragen zur Bewertungsqualität. Diese wurden jedoch nicht auf den Ausprägungsgrad hin kodiert (vgl. Anhang B). Dies verdeutlicht, dass die erste Lehrkraft im Unterrichtsfach Biologie, die Lehrkraft in Deutsch sowie die zweite Lehrkraft in Religion alle Bezugsnormen bei der Analyse textbasierter Schülerleistungen berücksichtigten. Die Bestimmung der Bezugsnorm auf Grundlage des Kodierleitfadens hatte jedoch ergeben, dass die beiden Kodierer unterschiedlicher Ansicht bei der Bestimmung der einzelnen Bezugsnormen waren. Offensichtlich ließen sich diese nicht deutlich genug aus den Interviewaussagen differenzieren.

Tabelle 94 Bezugsnormorientierung

		individuell		**sozial**		**kriterial**	
		1. MP	2. MP	1. MP	2. MP	1. MP	2. MP
Biologie	1. Lehrkraft	X		X	X	X	X
	2. Lehrkraft						
Deutsch		X	X	X	X	X	X
Religion	1. Lehrkraft		X				
	2. Lehrkraft	X	X		X	X	X
Kunst							X

8.6.4 Reliabilität induktiven Analysen

8.6.4.1 Reliabilitätskoeffizient (nach Krippendorf)

Zur Überprüfung der Zuverlässigkeit der Kategorienbildung hinsichtlich des induktiven Regelwerks wurde der Reliabilitätskoeffizient nach Krippendorf berechnet (vgl. Mayring, 2008 b, S.113; Bortz & Döring, 2003). Folgende Tabelle zeigt die Reliabilitätskoeffizienten in Bezug auf die Kategorienbildung sowie der Oberkategorienbildung. Im Unterrichtsfach Biologie ergab sich bei der Oberkategorienbildung eine hohe Messzuverlässigkeit. Die Kodierer bestimmten bei allen Textstellen dieselben Oberkategorien. Dies traf auf beide Messzeitpunkte zu bis auf das Nachinterview, welches mit der zweiten Lehrkraft geführt wurde. Die Kategorienbildung zeigte im Unterrichtsfach Biologie beim ersten Messzeitpunkt eine hohe Übereinstimmung und beim zweiten Messzeitpunkt eine gute Übereinstimmung. Das Nachinterview verdeutlichte eine geringe Übereinstimmung. Im Unterrichtsfach Deutsch zeigte sich bei beiden

Messzeitpunkten eine hohe Messzuverlässigkeit in Bezug auf die Kategorien- sowie der Oberkategorienbildung. Im Unterrichtsfach Religion zeigte sich bei der ersten Lehrkraft eine hohe Zuverlässigkeit der Oberkategorienbildung in Bezug auf den zweiten Messzeitpunkt. Bei der zweiten Lehrkraft im Unterrichtsfach Religion war die Zuverlässigkeit der Oberkategorienbildung zum ersten Messzeitpunkt hoch. Im Unterrichtsfach Kunst fand sich beim ersten Messzeitpunkt eine gute Zuverlässigkeit für die Kategorien- sowie der Oberkategorienbildung.

Tabelle 95 Reliabilitätskoeffizienten (nach Krippendorff) Bewertungskriterien

		Kategorie			**Oberkategorie**		
		1 MP	**2 MP**	**Nach-Interview**	**1 MP**	**2 MP**	**Nach-Interview**
		IR	IR		IR	IR	IR
Biologie	1. Lehrkraft	1.00	0.82		1.00	1.00	
	2. Lehrkraft	1.00	0.80	0.36	1.00	1.00	0.67
Deutsch		1.00	0.82		1.00	1.00	
Religion	1. Lehrkraft	0.00	0.55		0.00	1.00	
	2. Lehrkraft	0.57	0.59		1.00	0.50	
Kunst		0.86	0.00		0.86	0.00	

Die folgende Tabelle stellt die Reliabilitätskoeffizienten, hinsichtlich der induktiven Auswertung der Vorgehensweise der Lehrkräfte (von der Entwicklung der Klausur bis zur Bewertung) dar. Hierbei ergaben sich bei den Kategorien- sowie Oberkategorienbildungen niedrige bis keine Messzuverlässigkeiten.

Tabelle 96 Reliabilitätskoeffizienten (nach Krippendorff) Vorgehensweise

		Kategorie		**Oberkategorie**	
		1 MP	**2 MP**	**1 MP**	**2 MP**
		IR	IR	IR	IR
Biologie	1. Lehrkraft	0.67	0.00	0.00	0.00
	2. Lehrkraft	0.00	0.00	0.00	0.00
Deutsch		0.00	0.00	0.00	0.00
Religion	1. Lehrkraft	0.00	0.00	0.00	0.00
	2. Lehrkraft	0.00	0.27	0.00	0.67
Kunst		0.29	0.00	0.00	0.00

8.6.5 Reliabilität deduktive Analysen

8.6.5.1 Reliabilitätskoeffizient (nach Krippendorf)

Zur Überprüfung der Zuverlässigkeit der Kategorienbildung hinsichtlich des deduktiven Regelwerks wurde der Reliabilitätskoeffizient nach Krippendorf berechnet (vgl. Mayring, 2008 b, S.113; Bortz & Döring, 2003). Folgende Tabelle zeigt die Reliabilitätskoeffizienten. Dabei können die Werte des zweiten Messzeitpunktes im Unterrichtsfach Deutsch als hoch eingeschätzt werden (vgl. Bühner, 2004, S.129). Die induktiven Analysen der Interviewpassagen der ersten Lehrkraft im Unterrichtsfach Biologie verdeutlichten gute Messzuverlässigkeiten. Dies traf auch auf die zweite Lehrkraft im Unterrichtsfach Biologie zu, allerdings nur für den ersten Messzeitpunkt. Zudem zeigte sich für die zweite Lehrkraft im Unterrichtsfach Religion beim zweiten Messzeitpunkt und im Unterrichtsfach Deutsch beim ersten Messzeitpunkt eine gute Messzuverlässigkeit. Die Messgenauigkeit war in allen anderen Fällen niedrig.

Tabelle 97 Reliabilitätskoeffizienten (nach Krippendorff)

		1 MP	**2 MP**
		IR	IR
Biologie	1. Lehrkraft	0.87	0.89
	2. Lehrkraft	0.80	0.69
Deutsch		0.83	1.00
Religion	1. Lehrkraft	0.58	0.63
	2. Lehrkraft	0.72	0.90
Kunst		0.42	0.53

IR= Interkoderreliabilität

8.6.5.2 Interkoderreliabilität

Zur Überprüfung der Zuverlässigkeit der Ausprägungsgrade hinsichtlich des deduktiven Regelwerks wurde die Inderkoderreliabilität (nach Spearman) berechnet (vgl. Mayring, 2008 b, S.113; Bortz & Döring, 2003). Folgende Tabelle zeigt die Interkoderreliabilitäten. Dabei können die zusammengefassten Messzeitpunkte der ersten Lehrkraft im Unterrichtsfach Biologie als hoch eingeschätzt werden (vgl. Bühner, 2004, S.129). Alle anderen Werte ergaben eine niedrige Übereinstimmung der beiden Kodierer bezüglich der Ausprägungsgrade.

Tabelle 98 Rangkorrelationen der einzelnen MP‘S

		1 MP		**2 MP**		**1+2 MP**	
		IR	df	IR	df	IR	df
Biologie	1. Lehrkraft	0.13	11	0.63 (*)	11	0.94 (*)	24
	2. Lehrkraft	0.33	10	0.41	7	0.25	19
Deutsch		0.59 (*)	11	0.44	11	0.51	24
Religion	1. Lehrkraft	0.68 (*)	8	0.45	9	0.42 (*)	19
	2. Lehrkraft	0.80 (*)	12	0.61 (*)	12	0.68 (*)	26
Kunst		0.00	11	0.38	11	0.46 (*)	27

MP = Messzeitpunkt; IR = Interkoderreliabilität;

9 Diskussion der Hochschulergebnisse

Das zentrale Anliegen der vorliegenden Forschungsarbeit war die Überprüfung technologiegestützer Leistungsdiagnostik. Der empirische Forschungsstand zeigte bei der Bewertung von Lernergebnissen bei Lehrkräften eine unzureichende Berücksichtigung der Gütekriterien. Daraufhin wurde das in dieser Forschungsarbeit herangezogene Diagnoseinstrument in Erwägung gezogen und einer empirischen Überprüfung hinsichtlich einer möglichen Unterstützung für Lehrkräfte bei der Bewertung von Texten unterzogen. Dabei stand die Frage einer zeitökonomischen-, objektiven-, zuverlässigen- und gültigen Analyse textbasierter Lernergebnisse im Vordergrund. Im Hochschulkontext stand die mögliche Unterstützung inhaltlich orientierter Kriterienkataloge durch die semantischen Ähnlichkeitskennwerte durch T-MITOCAR im Vordergrund.

9.1 Diskussion der Ergebnisse der ersten Untersuchung

Die Betrachtung der herangezogenen Kriterien zeigte bei der ersten Aufgabenstellung des selbstregulierten Lernens eine gute bis hohe Auswertungsobjektivität. Das heißt, die Bewertungen ergaben auf Grundlage derselben Kriterien unabhängig vom Hochschuldozenten ähnliche Gesamteinschätzungen. Dies ergab gute Übereinstimmungen der Kriterien - bis auf die Kriterien der Definition und der Probleme und Grenzen des Modells. Die Messgenauigkeit der einzelnen Skalen war - nachdem die beiden nicht funktionierenden Kriterien entfernt wurden - gut. Es ergaben sich Alpha-Werte von $\alpha = 0.85$ (1. Bewerter) bzw. $\alpha = 0.80$ (2. Bewerter), d. h. die in dieser Untersuchung herangezogenen Bewertungskriterien entsprechen den angeforderten Reliabilitätskoeffizienten von $r > 0.80$ (vgl. Bühner, 2004, S. 129). Auch die Interkoderreliabilitätswerte waren bis auf das Kriterium der Definition gut bzw. hoch. Dies deutet auf eine hohe Ähnlichkeit der beiden Hochschuldozenten hinsichtlich der Bewertung auf Grundlage derselben Kriterien hin.

Die Auswertungsobjektivität der zweiten Aufgabenstellung bezüglich der Metakognition ergab niedrige Cohen's Kappa Werte. Die Untersuchung der Reliabilität mithilfe der Cronbach's Alpha-Werte ergab bei den Bewertungen des ersten Hochschuldozenten eine gute Messgenauigkeit ($\alpha = 0.87$). Dies traf auf die Bewertungen des zweiten Hochschuldozenten nicht zu ($\alpha = 0.79$). Die

Übereinstimmung beider Bewerter in Bezug auf die Gesamtskala war niedrig (r = 0.73). Demzufolge eigneten sich die Kriterien bei der Aufgabenstellung der Metakognition nicht. Zudem ergab die Betrachtung der individuellen Textlängen, dass die von den Studierenden externalisierten Texte bei dieser Aufgabenstellung deutlich kürzer ausfielen. Ein möglicher Grund hierfür könnte die Struktur des Lehrtextes gewesen sein deren erfragter Bereich im Vergleich zu dem Lehrtext der anderen Aufgabenstellung deutlich kürzer war. Dies wäre für weiter Forschungsfragen ein interessanter Aspekt, um aufzugreifen, inwiefern sich die Länge und Elaboriertheit des gelesenen Lehrtextes auf die mentale Modellbildung auswirkt.

Die Beantwortung der Frage, inwiefern sich die Leistungsbewertungen auf Grundlage inhaltlich orientierter Kriterien durch semantische Kennwerte abbilden lassen ergab, dass sich diese durch Concept Matching sowie Propositional Matching abbilden lassen. Dies war auf Grundlage der sozialen Bezugsnorm als Außenkriterium (Gesamtmodell) möglich. Obwohl die Software bezüglich der zugrunde gelegten Kriterien völlig blind war, war es möglich bei der Aufgabenstellung des selbstregulierten Lernens die Lernergebnisse mit einem Korrelationskoeffizienten (Spearman) von r = 0.58 und r = 0.61 (*) durch das Concept Matching vorherzusagen und mit einem Korrelationskoeffizienten (Spearman) von r = 37 und r = 0.34 (*) durch das Propositional Matching (vgl. Tabelle 16). Hier ergaben sich Korrelationskoeffizienten von r = 0.61 (Concept Matching) und r = 0.34 (Propositional Matching). Die Ergebnisse deuten somit darauf hin, dass die semantischen Kennwerte mögliche Hinweise geben um die Bewertung textbasierter Lernergebnisse vorherzusagen. Die hohen Korrelationskoeffizienten in Bezug auf die soziale Bezugsnorm können darin begründet werden, dass die Kriterien inhaltliche Schlagwörter enthielten, die in den einzelnen Texten enthalten sein sollten. Offensichtlich erhielten die Studierenden eine bessere Gesamtbewertung wenn sie Begrifflichkeiten ähnlich verwendeten wie die im Gesamtmodell enthaltenen. Bei der zweiten Aufgabenstellung war die semantische Abbildung der Kennwerte nicht möglich. Dies kann anhand der Kriterien begründet werden, die für diese Untersuchung herangezogen werden sowie an den kurzen Texten.

In der vorliegenden Untersuchung erfolgte - wie bereits in Kapitel 2.1.3 verdeutlicht - nicht eine Analyse des intern abgebildeten Wissens der an dieser

Teilstudie teilgenommenen Studierenden, sondern eine Analyse sowie Bewertung des in Textform externalisierten Wissens. Dies hat zur Konsequenz, dass die Lernenden möglicherweise viel mehr zu dem abgefragten Gegenstandsbereich wussten, als sie in den Texten abgebildet haben.
Die gute Auswertungsobjektivität der ersten Aufgabenstellung steht entgegen der Untersuchungen von Ingenkamp (1995). Jedoch wurde im Design der vorliegenden Studie keine Bezugsnormorientierung berücksichtigt. Es wurde nicht die Bewertung zwischen verschiedenen Lernergruppen verglichen. Demzufolge kann nicht ausgeschlossen werden, dass dies mögliche Effekte auf die Bewertung gehabt hätte. Auch mögliche Fehlerquellen im Prozess der Bewertung - wie beispielsweise einer gut leserlichen Handschrift wurden in der vorliegenden Arbeit nicht berücksichtigt.

9.2 Diskussion der Ergebnisse der zweiten Untersuchung

Die Betrachtung der herangezogenen Kriterien zeigte bei der ersten Aufgabenstellung des selbstregulierten Lernens eine niedrige bis gute Auswertungsobjektivität. Das heißt, beide Hochschuldozenten waren sich hinsichtlich der zugrunde gelegten Kriterien nicht einig, sondern sahen trotz gleicher Bewertungskriterien unterschiedliche Aspekte in den einzelnen Texten. Die Messgenauigkeit der einzelnen Skalen war bis auf die Gesamtskala niedrig. Der Alpha-Wert der Gesamtskala war jedoch bei beiden Hochschuldozenten hoch ($\alpha = 0.82$ beim ersten Hochschuldozenten und $\alpha = 0.87$ beim zweiten Hochschuldozenten). Das heißt, die in dieser Untersuchung herangezogenen Gesamtwerte entsprechen den angeforderten Reliabilitätskoeffizieten von $r > 0.80$ (vgl. Bühner, 2004, S. 129). Die Interkoderreliabilitätswerte waren bei allen Skalen niedrig. Dies deutet auf eine geringe Ähnlichkeit der beiden Hochschuldozenten hinsichtlich der Bewertung auf Grundlage derselben Kriterien hin.
Die Auswertungsobjektivität der zweiten Aufgabenstellung (Lernstrategien) ergab auf allen Kriterien bis auf Fragenstellen, Schreiben und Lernumgebung eine niedrige Auswertungsobjektivität. Die Untersuchung der Reliabilität zeigte bei beiden Hochschuldozenten eine gute Messgenauigkeit in Anbetracht der Gesamtskala (Lernstrategien) sowie der Gesamtskala Kognitive Lernstrategien. Die

Betrachtung der Interkoderreliabilitäten verdeutlichte bei allen Skalen eine gute bis hohe Übereinstimmung zwischen beiden Hochschuldozenten.

Die empirischen Befunde deuten darauf hin, dass sich die inhaltlich orientierten Bewertungen durch die semantischen Kennwerte abbilden lassen. Bei der ersten Aufgabenstellung des selbstregulierten Lernens konnte sich ein signifikanter Korrelationskoeffizient von $r = 0.23$ feststellen lassen (vgl. Tabelle 29). Bei der zweiten Aufgabenstellung der Lernstrategien zeigten sich signifikante Korrelationskoeffizienten von $r = 0.29$ (Propositional Matching) sowie $r = 0.25$ (Balanced Semantic Matching). Im Vergleich zur ersten Aufgabenstellung der ersten Hochschulstudie ließen sich die Bewertungen in dieser Untersuchung durch das Concept Matching deutlich geringer abbilden. Dies lässt sich durch die Kriterien erklären, die in dieser Studie geringere Objektivitäts- sowie Reliabilitätsmaße aufwiesen.

Die empirischen Befunde von Pirnay-Dummer (2012) zeigten, dass die semantischen und strukturellen Kennwerte keine Anteile der Notengebung im Fachbereich Erziehungswissenschaft abbildeten. In der vorliegenden Untersuchung konnten inhaltlich orientierte Anteile der Punktevergabe mittels semantischer Kennwerte abgebildet werden.

10 Diskussion der Schulergebnisse

Im Fokus der Schulstudien stand, neben der Überprüfung einer möglichen technologiegestützten Leistungsbewertung textbasierter Prüfungsleistungen, die Frage im Zentrum, inwiefern sich die Bewertungskriterien von Lehrkräften sowie deren textbasierter Erwartungshorizont (Musterlösung) durch eine gezielte Intervention verändern lassen.

10.1 Diskussion der Untersuchung im Fach Biologie

Die empirischen Befunde der vorliegenden Untersuchung zeigten eine niedrige Bewertungsübereinstimmung (vgl. Tabelle 39) zwischen den Lehrkräften. Dies lässt darauf schließen, dass die Kriterien die durch die Lehrkräfte vergebenen Leistungspunkte beeinflussen. Für den Schüler entscheidet in dem Fall der Kriterienkatalog, ob seine Prüfungsleistung als eher gut oder mittelmäßig bewertet wird. Der zweite Kriterienkatalog führte zu deutlich stabileren Einschätzungen zwischen den Messzeitpunkten (vgl. Tabelle 40). Beide Lehrkräfte verwendeten unterschiedliche Bewertungskriterien. Die Untersuchung der strukturell orientierten Bewertungskriterien zeigte, dass beide Lehrkräfte Aspekte hinsichtlich des Layouts bewerteten (beispielsweise, ob der Text gut strukturiert war und ob die Fachtermini richtig verwendet wurden). Die Bewertungskriterien der Lehrkräfte umfassten inhaltliche Aspekte, die von den Schülern angesprochen werden konnten oder nicht. Die induktive Auswertung der Interviews ergab, dass beide Lehrkräfte angaben, sich an die Kompetenzbereiche (Wiedergabe, Verständnis sowie Transfer) anzulehnen. Beide Bewertungskriterien zeigten vor allem ein Abbilden von deklarativem Wissen, sodass die Schüler ihr Wissen wiedergaben. Die Kompetenzpunkte der zweiten Lehrkraft innerhalb des Kriterienkatalogs (siehe Tabelle 35) ließen Verständnis und Transfer nicht erkennen. Die Analyse des Kriterienkatalogs der zweiten Lehrkraft sowie deren Anwendung im Bewertungsprozess zeigte, dass in den Schülertexten primär nach den im Kriterienkatalog angegebenen Begrifflichkeiten gesucht wurde (vgl. Tabelle 80). Die erste Lehrkraft im Unterrichtsfach Biologie gab an, bei der Materialnutzung sowie der Antwort der Schüler auf einen schlussfolgernden Bezug, sowie eine Begründung der eigenen Stellungnahme, zu achten (vgl. Tabelle 78, Zeile: 25-38). Dies verdeutlicht, dass mithilfe der Kriterien über das reine Wiedergeben von

Wissen hinaus bewertet wurde. Die Analyse der induktiven Auswertung in Bezug auf die Vorgehensweisen lässt bei der ersten Lehrkraft auf eine hohe diagnostische Expertise schließen, was die deduktiven Analysen bestätigten. Diese Lehrkraft berücksichtigte die vom Lehrplan vorgegebenen Inhalte im Unterricht und orientierte sich bei den Klausurinhalten an den Unterrichtsinhalten, was auf eine hohe Inhaltsvalidität der Klausurfragen schließen lässt. Zudem berücksichtigte diese Lehrkraft unterschiedliche Schwierigkeitsgrade. Die deduktive Auswertung zeigte eine hohe Bewertungsqualität bei der ersten Lehrkraft und eine mittelmäßige Bewertungsqualität bei der zweiten Lehrhaft. Dies verdeutlicht, dass die erste Lehrkraft stark die Gütekriterien berücksichtigte sowie Bewertungsfehler im Prozess zu minimieren suchte und geeignete Maßnahmen hierfür explizit nannte. Der Ausprägungsgrad der Bewertungsqualität änderte sich nach der Intervention nicht. Demzufolge verringerte sich der Ausprägungsgrad in Bezug auf die Bewertungsqualität bei der ersten Lehrkraft durch die Maßnahme nicht. Da dieser jedoch bei der zweiten Lehrkraft nicht gesteigert werden konnte, benötigt dies demzufolge eine gezieltere und umfassendere Schulung um das Konstrukt der Bewertungsqualität zu verändern. Die Lehrkräfte änderten beide ihre Kriterien nach der Intervention nicht. Die Bewertungskonstrukte der Lehrkräfte waren offensichtlich so stabil (möglicherweise bereits Schematisiert), dass sie eine andere Maßnahme benötigten, um sich verändern zu lassen. Auch die Musterlösung wurde von der ersten Lehrkraft nicht verändert. Da diese erst nach Abschließen der zweiten Bewertungsphase abgegeben wurde, kann nicht ausgeschlossen werden, dass in der ersten Bewertungsphase keine textbasierte Musterlösung erstellt worden war.

Dass die zweite Lehrkraft keine textbasierte Musterlösung erstellte, kann zudem auf die ungewohnte Situation zurückzuführen sein. Die Begründung, dass sich hinsichtlich der Aufgabenstellung keine textbasierte Ideallösung ausformulieren ließ, konnte durch die erste Lehrkraft nicht bestätigt werden. Die inhaltlichen sowie strukturellen Kriterien ließen sich weder durch die strukturellen noch durch die semantischen Kennwerte abbilden, was sich an der Musterlösung begründen lässt. Diese enthielt teilweise nicht ausformulierte Begrifflichkeiten, was einen Vergleich mit T-MITOCAR stark beschränkt. Zudem zeigte sich, dass nicht alle in den Bewertungskriterien enthaltenen Kriterien in der Musterlösung abgebildet waren,

was zudem einen inhaltsvaliden Vergleich des Außenkriteriums mit den einzelnen Schülerleistungen erschwert.
Die Post-Hoc-Analyse zeigte, dass die Textlänge bei der ersten Lehrkraft Einfluss auf die Bewertung hatte (vgl. Tabelle 42). Nach der Intervention fand sich dieser Einfluss noch - war jedoch geringer.
Aus den Befunden der vorliegenden Untersuchung zeigen die Bewertungen der Lehrkräfte unterschiedliche Herangehensweisen der Analyse von Texten im Vergleich zu T-MITOCAR. Demzufolge benötigen Lehrkräfte eine andere Unterstützung hinsichtlich der Analyse textbasierter Schülerleistungen. Das beinhaltet zunächst eine Optimierung der Kriterien, sodass diese explizit ausformuliert werden. Zudem bedarf es weiterer Innovation, um die zugrunde gelegten Kriterien in den jeweiligen Referenzmodellen abzubilden.
Die vorliegende Untersuchung umfasst nur zwei Lehrkräfte im Unterrichtsfach Biologie. Deswegen können keine systematischen Schlussfolgerungen getroffen werden. Die Ergebnisse zeigen, dass, wenn Lehrkräfte aufgefordert werden, Bewertungskriterien zu erstellen, sie Schwierigkeiten haben diese explizit zu nennen. Demzufolge scheiterte die Anwendung der kriterialen Bezugsnorm schon früher als erwartet.

10.2 Diskussion der Untersuchung im Fach Deutsch

Die empirische Überprüfung in der vorliegenden Untersuchung ergab eine niedrige Bewertungsübereinstimmung bei wiederholter Bewertung der textbasierten Schülererörterungen (vgl. Tabelle 47). Dies bestätigt die Befunde von Hartog & Rhodes (1936, 1995), Dicker (1977, 1995) und Aschersleben (1971), dass die Messzuverlässigkeit klassischer schriftlicher Prüfungssituationen unzureichend ist.
Die Bewertungskriterien, die von der Lehrkraft entwickelt und für die Bewertung herangezogen wurden, zeigten einen relativ offenen Erwartungshorizont, welcher die Schüler in keine Richtung hinsichtlich ihrer Argumentationsweisen zwang. Die inhaltlich orientierten Kriterien (vgl. Tabelle 44) spezifizierten die formal-orientierten Kriterien in Bezug auf eine erkennbare dreigliedrige Struktur. Sie umfassten keine impliziten inhaltlichen Kriterien die im Text zu finden sein mussten, sondern strukturelle Aspekte, wie beispielsweise ob die Argumente sinnvoll angeordnet waren und Überleitungssätze vorhanden waren. Der

signifikante Korrelationskoeffiziente (nach Spearman) hinsichtlich der strukturellen Ähnlichkeitsmaße (Gamma beim ersten Messzeitpunkt) lässt sich dadurch erklären. Der (signifikant) negative Zusammenhang von $r = -0.42$ verdeutlicht, dass, je ähnlicher die Begrifflichkeiten in den einzelnen Schülertexten im Vergleich zu den textbasierten Außenkriterien vernetzt wurden, desto besser war die erzielte Note (vgl. Tabelle 49). Dies betraf sowohl die kriteriale Bezugsnorm als auch die soziale Bezugsnorm.

Die induktiven Analysen bezüglich der Kriterien zeigten, dass die in den Bewertungskriterien inhaltlich orientierten Kriterien Aspekte hinsichtlich der strukturellen Anordnung des Textes umfassten. Dies erklärt, wieso der strukturelle Kennwert (Gamma) signifikant mit der Gesamtnote korrelierte.

Die induktiven Analysen hinsichtlich der Vorgehensweise der Lehrkraft von der Klausurerstellung bis hin zur Bewertung ergaben bei beiden Messzeitpunkten eine Orientierung an den Unterrichtsinhalten. Die Lehrkraft änderte die Kriterien nach der Intervention in der Gewichtung. Vor der Intervention wurden die einzelnen Kriterien mehr zur Orientierung herangezogen, um eine Notentendenz zu erkennen. Nach der Intervention wurden die einzelnen Kriterien fünfstufig gewichtet und so die Gesamtnote ermittelt. Dies weist auf eine Präzision der Ermittlung der Gesamtbewertung auf Grundlage des Erwartungshorizontes hin. Nach der Intervention erfolgte keine inhaltliche Änderung der einzelnen Kriterien. Daran erkennt man, dass dieses Konstrukt relativ stabil ist. Es konnte keine Veränderung der Musterlösung festgestellt werden, da diese erst in der zweiten Bewertungsphase erstellt worden war. Deswegen können mögliche Effekte des Erstellens der Musterlösung in der zweiten Bewertungsphase in Bezug auf die Einschätzung der Schülerleistungen nicht ausgeschlossen werden.

Die deduktiven Analysen in Hinsicht auf die Bewertungsqualität der Lehrkraft zeigten zu beiden Messzeitpunkten eine mittlere bis hohe Ausprägung. Da jedoch die Interkoderreliabilität beim zweiten Messzeitpunkt nicht signifikant war, können mögliche Schlüsse auf die eigentliche Bewertungsqualität dieser Lehrkraft auf Grundlage der Interviews nur mit äußerster Vorsicht gezogen werden. Möglicherweise waren die Kodierregeln für die Ausprägungsgrade nicht präzise genug.

Die Post-Hoc-Untersuchung zeigte vor und nach der Schulung keinen Einfluss der Textlänge auf die Bewertung. Das lässt vermuten, dass sich die Lehrkraft bei der

Bewertung nicht von der Textlänge beeinflussen ließ. Dies spricht gegen die empirischen Befunde von Birkel & Birkel (2002).
Die vorliegenden Ergebnisse verdeutlichen, dass im Unterrichtsfach Deutsch bei der Bewertung von Aufsätzen wo Aspekte der Struktur und Anordnung eine wesentliche Rolle spielen, die strukturellen Kennwerte einen möglichen Indiz in Bezug auf die strukturelle Textqualität geben. Dies trifft die Elaboriertheit von Texten im Vergleich zu den Außenkriterien. Da das inhaltliche Antwortspektrum bei Aufsätzen ziemlich weit ist, können die semantischen Kennwerte auf Grundlage der vorliegenden Ergebnisse nicht als Indiz für die inhaltliche Qualität von Texten herangezogen werden. Dies scheitert schon deswegen, weil sich für Aufsätze inhaltliche Kriterien nur schwer präzisieren lassen.
Die vorliegende Studie umfasste nur eine Lehrkraft, deswegen können keine systematischen Schlussfolgerungen auf weitere Lehrkräfte im Unterrichtsfach Deutsch getroffen werden.

10.3 Diskussion der Untersuchung im Fach Religion

Beide Lehrkräfte zeigten niedrige Werte hinsichtlich der Bewertungsstabilität zwischen den Messzeitpunkt. Das heißt, es kann nicht ausgeschlossen werden, dass der Zeitpunkt, zu dem die Schülerleistung bewertet wird, sich auf die Bewertung auswirkt. Offensichtlich waren die Kriterien zu unpräzise, um bei wiederholter Bewertung dieselben Leistungspunkte zu erzielen. Die in den beiden Teilstudien erzielten Lehrerurteile sind nicht stabil genug, um den wissenschaftlichen Standards zu genügen. Beide Lehrkräfte änderten ihre gemeinsam erstellte Musterlösung und Kriterienkataloge nach der Intervention nicht. Die zweite Lehrkraft änderte die Gewichtungen der inhaltlichen und strukturellen Bewertung, sodass beim zweiten Messzeitpunkt die strukturellen Aspekte nur noch mit 20 % in die Note mit einflossen, anstatt, wie beim ersten Messzeitpunkt mit 30 %.
Ein Grund zur Annahme, dass Bewertungskonstrukte von Lehrkräften zu stabil sind, um durch eine einmalige Maßnahme verändert zu werden. Die Gesamtbewertung der ersten Lehrkraft ließ sich weder durch die strukturellen noch durch die semantischen Kennwerte abbilden. Ein möglicher Grund, dass sich die inhaltlichen Kriterien nicht abbilden ließen, könnte gewesen sein, dass hinsichtlich der dritten Aufgabenstellung keine textbasierte Musterlösung erstellt wurde, da

diese bei der zweiten Lehrkraft aus der Bewertung entfernt wurde. Die erste Lehrkraft hatte dennoch Punkte für diese Aufgabenstellung gegeben. Dass sich die strukturellen Kennwerte bei der ersten Lehrkraft nicht abbilden ließen, lässt sich an den Kriterien begründen. Hier ergab sich den Interviewaussagen zufolge ein Abzug (von insgesamt einem Punk) bei nicht Einhalten z. B. der Grammatikregeln. Die Gesamtbewertung der zweiten Lehrkraft ließ sich vor der Intervention durch den strukturellen Kennwert Surface auf Grundlage des Gesamtmodells (der sozialen Bezugsnorm) abbilden. Das Ergebnis lässt vermuten, dass die Lehrkraft sich bei der Bewertung an der sozialen Bezugsnorm orientierte, was sich in den induktiven Analysen bei der Vorgehensweise bestätigte. Hier gab die Lehrkraft an, den Erwartungshorizont, falls notwendig, im Nachhinein an die Klasse anzupassen (vgl. Tabelle 90, Zeile: 133 - 150). Dass sich die strukturelle Bewertung der zweiten Lehrkraft beim zweiten Messzeitpunkt nicht mehr abbilden ließ, könnte auf die Verringerung der strukturellen Gewichtung bei der Gesamtbewertung zu schließen sein.

Da die erste Lehrkraft die dritte Aufgabenstellung mit in die Gesamtbewertung fließen ließ, obwohl diese nicht Bestandteil der textbasierten Musterlösung war, kann hier nicht auf einen inhaltsvaliden Vergleich der Schülerleistungen hinsichtlich der Musterlösung geschlossen werden. Das erklärt, warum sich möglicherweise die Bewertungen der ersten Lehrkraft nicht durch die Kennwerte abbilden ließen. Da in der ersten Aufgabenstellung sprachlich orientierte Aspekte gegeben waren, begründet dies zudem, warum ein semantischer Vergleich durch die inhaltlich orientierten Kriterien der ersten Lehrkraft nicht abgebildet werden konnten. Für weitere Untersuchungen müssten diese deutlicher getrennt werden.

Die induktiven Analysen hinsichtlich der Vorgehensweisen zeigten, dass die erste Lehrkraft bei der Bewertung der dritten Aufgabenstellung, welche Transfer erfasste, Kreativität bewertete. Das lässt eine geringe Auswertungsobjektivität vermuten, da Kreativität ein schwer messbares Konstrukt darstellt. Die erste Lehrkraft gab an, sich auf Grundlage von „Baucheinschätzung“ einen ersten Eindruck der einzelnen Schülerleistungen zu verschaffen. Dies deutet auf eine geringe Auswertungsobjektivität hin. Die deduktiven Analysen ergaben bei beiden Lehrkräften vor und nach der Intervention eine mittelmäßige Bewertungsqualität. Diese ließen sich durch die Schulung nicht optimieren. Offensichtlich sind

gezieltere Schulungen notwendig, um die Bewertungsqualität von Lehrkräften zu steigern.

Die Post-Hoc-Untersuchung zeigte, dass sich beide Lehrkräfte bei der Bewertung von der Textlänge beeinflussen lassen. Dies bestätigt die empirischen Befunde von Birkel & Birkel (2002). Dieser Effekt verringerte sich nach der Schulung bei beiden Lehrkräften. Die vorliegenden Teilstudien im Unterrichtsfach Religion zeigten, dass Lehrkräfte trotz gleicher Bewertungskriterien unterschiedliche Gewichtungen vornehmen im Zusammenhang mit der Notengebung., während die erste Lehrkraft für die generelle Bewertung nur einen Leistungspunkt Abzug bei Nichteinhalten der Grammatikregeln vergab, flossen bei der zweiten Lehrkraft 30 % (vor der Intervention) bzw. 20 % (nach der Intervention) mit Hinsicht auf die generelle Bewertung in die Gesamtnote ein. Dieser Unterschied zeigt, dass Schülerleistungen unterschiedliche Leistungspunkte erzielen, auch wenn dieselben Kriterien angelegt werden. Die Befunde ergaben erste Hinweise dafür, dass sich die Bewertungen im Unterrichtsfach Religion durch strukturelle Kennwerte abbilden lassen. Das heißt, dass auf Grundlage der herangezogenen Musterlösungen - welche die Kriterien der zweiten Lehrkraft vollständig abbildeten - unterschiedliche Expertisegrade ausfindig gemacht werden können.

10.4 Diskussion der Untersuchung im Fach Kunst

Die Beurteilungsstabilität der an dieser Untersuchung teilgenommen Lehrkraft zeigte eine niedrige Übereinstimmung der Gesamtbewertungen hinsichtlich der textbasierten Prüfungsleistungen der einzelnen Schüler bei wiederholter Bewertung (vgl. Tabelle 74). Dieses Ergebnis gibt Grund zur Annahme, dass die von der Lehrkraft entwickelten Kriterien nicht den wissenschaftlichen Standards entsprachen. Die **induktiven Analysen** der Bewertungskriterien verdeutlichten Schwierigkeiten im Erstellen valider Kriterien bezüglich der Interpretation der Bildbetrachtung. Dieses Kriterium umfasste Aspekte wie Eigenständigkeit, Aufbau sowie Originalität. Dabei gab die Lehrkraft an, in diesem Bereich individuell zu bewerten und verwies auf die Unmöglichkeit diesen Bereich explizit zu formulieren. Eine Post-Hoc-Analyse der Güte der durch die Lehrkraft erstellten textbasierten Musterlösung zeigte, dass diese Kriterien hinsichtlich der Interpretation nicht in der Musterlösung abgebildet waren. Da ein Großteil der

Bewertung in die Interpretation floss, könnte hier ein möglicher Grund für die geringe Bewertungsübereinstimmung bei wiederholter Messung zu finden sein. Die Kriterien wurden nach der Intervention nicht verändert, ebenso die Musterlösung, was darauf hindeutet, dass sich die expliziten Konstrukte der Bewertung nicht so einfach durch eine einmalige Schulung verändern lassen. Die von der Lehrkraft erstellten Kriterien ließen sich weder durch die semantischen noch durch die strukturellen Kennwerte abbilden. Dass die strukturellen Kennwerte sich durch die Gesamtbewertung der Lehrkraft nicht abbilden ließen, lässt sich auf Grundlage der Kriterien erklären. Diese enthielten keine strukturellen Komponenten - bis auf den Aufbau der Arbeit (vgl. Tabellen 70 - 71). In den Interviews gab die Lehrkraft hier an, dass es bei diesem um Aspekte der Nachvollziehbarkeit ginge. Somit enthalten die Kriterien keine strukturellen Aspekte, wie sie durch die strukturellen Kennwerte abgebildet werden. Eine Post-Hoc-Analyse zur Untersuchung inwiefern die explizierten Kriterien auch in der von der Lehrkraft festgestellten Musterlösung abgebildet waren, hatte ergeben, dass nicht alle in den Kriterien enthaltenen Aspekte in der Musterlösung abgebildet waren. Da diese allerdings als textbasiertes Außenkriterum für den Vergleich herangezogen wurde, ist nicht auszuschließen, dass sich die inhaltlichen Kriterien nicht durch die semantischen Kennwerte abbilden ließen, weil einzelne inhaltliche Kriterien in der Musterlösung fehlten. Die induktiven Analysen der Vorgehensweise der Lehrkraft ergaben, dass es der Lehrkraft unmöglich war den idealen Erwartungshorizont zu formulieren und dass diese deswegen individuell bewertete. Dies betraf die dritte Aufgabenstellung bezüglich der Interpretation. Da jedoch ein Schwerpunkt der Punkteverteilung in die Interpretation floss (vgl. Tabelle 90), führte dies notwendigerweise zu großen Verzerrungen bei der Bewertung. Die Befunde zeigen, dass Lehrkräfte insbesondere bei der Analyse von Textbestandteilen hinsichtlich der Interpretation (Eigenständigkeit, Aufbau, Originalität) geschult werden sollten. Da Eigenständigkeit und Originalität sich nicht durch die Ähnlichkeitsmaße bestimmen lassen, bedarf es hier einer völlig anderen Unterstützung für Lehrkräfte als zunächst vermutet. Diese setzt nicht erst im Prozess der Analyse von Texten ein, sondern bereits viel früher - nämlich beim Operationalisieren und Messbar machen von inhaltsvaliden Kriterien.

Die deduktiven Analysen ergaben eine hohe Bewertungsqualität. Dies deutet auf eine starke Berücksichtigung der Gütekriterien sowie der Vermeidung von

Fehlerquellen im Prozess der Bewertung hin. Da die Bestimmung der Kategorien sowie die Bestimmung der Ausprägungsgrade jedoch eine niedrige Übereinstimmung zwischen den Kodierern aufwies, können diese Schlüsse hinsichtlich der Bewertungsqualität nur mit äußerster Vorsicht gezogen werden.

10.5 Einfluss von Fachstruktur und Fachlogik

Im *naturwissenschaftlichen* Fach war zu erwarten, dass sich vor allem Kriterien der richtigen Anwendung und Einordnung von Fachbegriffen vorfanden sowie konkrete Beispiele. Die Verwendung von Beispielen fand sich in den Kriterien vor. Dies war jedoch in der Aufgabenstellung begründet, da die Schüler aufgefordert waren eigene Beispiele zu finden. Die Aufgabenstellung zeigte, dass die Schüler wie in einer Erörterung aufgefordert wurden, ihre eigene Meinung zu begründen. Dies wäre eher in einem sprachlich orientierten Fach wie Deutsch zu erwarten gewesen.

Im *sprachlichen* Fach (Deutsch), war zu vermuten, dass sich in den Kriterien vor allem sprachliche Aspekte einer guten Ausdrucksweise vorfinden. Die Ergebnisse zeigten, dass sich hier vor allem Kriterien fanden, welche die Struktur des Textes betrafen.

Im *gesellschaftlichen* Fach Religion war anzunehmen, dass sich vor allem Kriterien vorfinden, welche das Begründen der eigenen Stellungnahme berücksichtigen. Diese fanden sich in den Bewertungskriterien vor.

Im *musisch-künstlerischen* Fach waren aus fachlogischer Perspektive vor allem Kriterien zu vermuten, welche die Kreativität der Schüler berücksichtigte. Dieses fand sich in der Originalität, welches durch das Stuttgarter Modell festgelegt war. Allerdings blieb unklar, wie dieses genau gewichtet und bewertet wurde. Hier gab die Lehrkraft an, individuell zu bewerten und zu gewichten. Die induktiven Analysen ergaben, dass die Lehrkraft das flexible Einbringen neuer Gedanken berücksichtigte.

10.6 Methodenkritik

Der folgende Abschnitt fokussiert eine kritische Betrachtung der methodischen Vorgehensweise im Kontext der Untersuchungen des Schulkontextes. Dabei erfolgt eine kritische Auseinandersetzung der quantitativen wie auch der qualitativen

Erhebungen und der jeweiligen Auswertungsverfahren. Schließlich werden die textbasierten Musterlösungen als Außenkriterium kritisch betrachtet.
Bei der Transkription der Klausuren kann es möglicherweise in einzelnen Fällen zu Übertragungsfehlern gekommen sein. Manche Schriftbilder waren schwer zu lesen. Dies hat jedoch für den Vergleich mit T-MITOCAR nur geringe Auswirkungen, da hier nur die Nomina der 30 stärksten Modellrelationen mit im Modell enthalten sind.
Im Unterrichtsfach Biologie war die zweite Lehrkraft nicht am Unterrichtsgeschehen beteiligt. Deswegen kann hier nicht ausgeschlossen werden, dass inhaltliche Kriterien abgebildet wurden, die nicht im Unterricht berücksichtigt worden waren. Bei der Transkription der Interviews waren kleine Auszüge nicht deutlich zu verstehen. Da dies jedoch nur wenige Passagen betraf, kann eine Verzerrung der Ergebnisse ausgeschlossen werden. Das erste Interview (mit der ersten Lehrkraft im Unterrichtsfache Biologie) wurde aus technischen Gründen nicht aufgenommen. Dies wurde nach der Methode der Rekonstruktion unmittelbar nach dem geführten Interview von der Person aufgeschrieben, die das Interview geführt hatte. Hier kann es möglicherweise zu Verzerrungen gekommen sein. Auch kann nicht ausgeschlossen werden, dass für das Interview wichtige inhaltliche Aspekte nicht mehr in dem Umfang erinnert werden konnten, wie sie von der Lehrperson gesagt worden waren. Für die Ergebnispräsentation wurde aus Reliabilitätsgründen auf die Interpretation nicht gleich kodierter Interviewauszüge verzichtet. Dies betraf die deduktiven Analysen. Nur die durch beide Kodierer als gleich eingeschätzten Kategorien wurden betrachtet. Dies kann möglicherweise dazu geführt haben, dass für die Auswertung wichtige Aspekte wegfielen. Für die Bestimmung der Messzuverlässigkeit der Oberkategorien der induktiven Analysen wurden nur die gemeinsamen Oberkategorien herangezogen, deren Kategorien bereits gleich bestimmt worden waren. Dies ist eine starke Einschränkung und kann dazu geführt haben, dass die Messzuverlässigkeit für die Oberkategorien gering ausfiel.

11 Zusammenfassung

In der vorliegenden Arbeit stand die Überprüfung einer möglichen technologiegestützten Leistungsfeststellung im Vordergrund. Diese sollte eine zeitökonomische-, objektive-, reliable- und valide Leistungsdiagnose berücksichtigen. Da insbesondere in Bildungsinstitutionen wie der Hochschule und der Schule selektive Funktionen obliegen, welche die Bewertung textbasierter Prüfungsleistungen verlangt, lag ein besonderes Augenmerk in diesen beiden Bereichen. In den Hochschulstudien stand zunächst die Frage im Raum, ob sich die inhaltlich orientierte Bewertung von Klausurergebnissen durch im Kontext der mentalen Modellbildung entwickelte Technologie abbilden lassen. In den Schulstudien interessierte zudem, ob sich die Bewertungskriterien und ausformulierten Erwartungshorizonte (Musterlösungen) durch eine gezielte Intervention in Bezug verändern lassen.

11.1 Zusammenfassung der Hochschulergebnisse

Die empirischen Befunde im Hochschulkontext geben erste Hinweise darauf, dass sich inhaltlich orientierte Leistungsbewertung durch technologiegestützte Kennwerte abbilden lassen, obwohl das Instrument völlig blind war, was die zugrunde gelegten Bewertungskriterien anbelangt. Dies gelang auf Grundlage der sozialen Bezugsnorm und der kriterialen Bezugsnorm. Am besten eignete sich dabei das Concept Matching, was in den herangezogenen Kriterien begründet ist, die Begrifflichkeiten umfassten, die in den einzelnen Prüfungsleistungen enthalten sein sollten. Da sich die durch die Hochschuldozenten vergebenen Gesamtpunkte jedoch nicht in allen Fällen durch die semantischen Kennwerte abbilden ließen, müssen diese Befunde kritisch betrachtet werden.

Zum Schluss deuten die Befunde auf eine erste zeitökonomische Rückmeldung der festgestellten textbasierten Leistungen hin, sobald ein geeignetes Außenkriterium gefunden ist. Bislang ermöglicht die in dieser Studie herangezogene Technologie es jedoch nicht, sprachstilistische Aspekte wie Ausdruck, richtige Verwendung der Grammatikregeln usw. zu berücksichtigen. Für eine konkrete Umsetzung im Bereich der Leistungsermittlung müsste die vorliegende Technologie erweitert und hinsichtlich der Notengebung spezifiziert werden, sodass dem Lerner auf

Grundlage der festgestellten Kennwerte rückgemeldet werden kann, ob sich seine Leistung z. B. in einem sehr guten oder einem ausreichenden Bereich befindet.

11.2 Zusammenfassung der Schulergebnisse

Die empirischen Befunde der vorliegenden Untersuchungen im Schulkontext deuten darauf hin, dass sich die strukturell orientierte Textanalyse durch strukturelle Kennwerte abbilden lässt. Diese Erkenntnis ließ sich jedoch nicht in den Fächern Biologie und Kunst bestätigen, da sich hier herausstellte, dass entweder die Kriterien etwaas anderes abbildeten als die Musterlösungen oder die Kriterien in Bezug auf einzelne Kriterien (wie beispielsweise die Interpretation der Bildbetrachtung in Kunst) sicht operationalisiert waren. Außerdem konnte gezeigt werden, dass Lehrkräfte zunächst eine völlig andere Unterstützung benötigen als zunächst vermutet. Diese setzt bereits im Prozess des Erstellens und Operationalisierens der einzelnen Kriterien an und nicht erst bei der Analyse von Texten. Zudem deuten die Ergebnisse auf eine geringe Bewertungsstabilität und somit auf eine geringe Zuverlässigkeit der angelegten Kriterien bei allen an diesen Untersuchungen teilgenommenen Lehrkräften hin. Demzufolge reicht es nicht, anzunehmen, dass die diagnostische Expertise reiche um eine objektive-, reliable- sowie valide Leistungsfeststellung zu erzielen. Um dies zu erreichen muss zunächst die Entwicklung und Optimierung der Kriterien ins Visier genommen werden. Die Untersuchungen zeigten, dass sich vor allem Kriterien welche die Analyse von Textaufbau und Struktur untersuchen möglicherweise durch die in diesen Studien herangezogene Technologie unterstützen lassen. Obwohl die Technologien blind für die Kriterien waren, konnten diese zu einem Großteil abgebildet werden. Für einen praktikablen Einsatz im Schulalltag wäre es interessant aus den Ähnlichkeitskennwerten mögliche Notenvorschläge zu generieren. Für eine präzisere Rückmeldung an den Lernenden wäre eine Berücksichtigung der einzelnen - in den Kriterien vorgegebenen - Gewichtungen sinnvoll.

Der Mixed-Methods-Ansatz ermöglichte es, einen tieferen Einblick in die durch die Lehrkräfte entwickelten Bewertungskriterien sowie deren konkreten Bewertung zu erzielen. Ebenso konnten die Notengebungen der Lehrkräfte mittels der Analyse der zugrunde gelegten Bewertungsqualität fundierter betrachtet und eingeordnet werden.

Literaturverzeichnis

Aebli, H. (1981). *Denken: Das Ordnen des Tuns.* Band 2: Denkprozesse. Stuttgart: Klett-Cotta.

American Psychology Association. (2007). *Publication manual of the American Psychology Association.* Washington, DC: American Psychological Association.

Beckenkamp, M. (1995), *Wissenspsychologie : zur Methodologie kognitionswissenschaftlicher Ansätze* , Asanger , Heidelberg .

Besser, M., & Krauss, S. (2009). Zur Professionalität als Expertise. In O. Zlatkin-Troitschanskaia, K. Beck, D. Sembill, R. Nickolaus & R. Mulder (Eds.), *Lehrprofessionalität – Bedingungen, Genese, Wirkungen und ihre Messung* (S. 71-82). Weinheim: Beltz.

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

Bohl, T. (2005). *Prüfen und Bewerten im Offenen Unterricht.* Weinheim und Basel: Beltz.

Boekaerts, M. (1999). Self-regulated learning. *International Journal of Educational Research, 31*, 445-457.

Bortz, J. & Döring, N. (2003). *Forschungsmethoden und Evaluation.* Berlin, Heidelberg, New York: Springer.

Boud, D., & Falchikov, N. (1995). What does research tell us about self assessment? In D. Boud (Ed.), *Enhancing learning through self assessment* (S. 115-166). London: Kogan Page.

Bromme, R. (1992). *Der Lehrer als Experte. Zur Psychologie des professionellen Wissens.* Bern: Huber.

Bromme, R. (2008). Lehrerexpertise. Teacher's Skill. In W. Schneider & M. Hasselhorn (Eds.), *Pädagogische Psychologie. Ein Lehrbuch* (S. 269-356). Weinheim: Beltz.

Bromme, R. & Haag, L. (2004). Forschung zur Lehrerpersönlichkeit. In W. Helsper & J. Böhme (Eds.), *Handbuch der Schulforschung* (S. 777-793). Wiesbaden: Verlag für Sozialwissenschaften.

Bromme, R., Rheinberg, F., Minsel, B., Winteler, A., Weidenmann, B. (2006). Die Erziehenden und Lehrenden. In A. Krapp & B. Weidenmann (Eds.), *Pädagogische Psychologie. Ein Lehrbuch* (S. 269-356). Weinheim: Beltz.

Bruner, J. S. (1964). The course of cognitive growth. *American Psychologist, 19*, 1-16.

Bryant, A. & Charmaz, K. (2007). The Sage Handbook of Grounded Theory. London: Sage

Bühner, M. (2004). *Einführung in die Test- und Fragebogenkonstruktion*. München: Pearson Studium.

Clariana, R. B. (2004). *ALA-Reader software,* version 1.01. Retrieved December 24, 2008, from http://www.personal.psu.edu/rbc4/score.htm

Clariana, R. B. (2010). Deriving Individual and Group Knowledge Structure from Network Diagrams and from Essays. In D. Ifenthaler, P. Pirnay-Dummer & N. M. Seel (Eds*.), Computer-Based Diagnostics and Systematic Analysis of Knowledge* (S. 117-130). New York: Springer.

Couné, B., Hanke, U., Ifenthaler, D. & Seel, N. M. (2003). *Modellkonstruktionen beim Problemlösen im Kontext entdeckenden Lernens: Eine Studie zur Implementierung theoretisch-begründeter Instruktionsprinzipien. Erster Bericht aus dem Forschungsprojekt „Modell-begründetes Lernen und Lehren. Multimediale Lernumgebungen als Gelegenheiten zum Nachdenken*. Freiburg: Institut für Erziehungswissenschaft.

Couné, B., Hanke, U., Ifenthaler, D. & Seel, N. M. (2004). *Modellkonstruktionen beim Problemlösen im Kontext entdeckenden Lernens: Eine Studie zur Implementierung theoretisch-begründeter Instruktionsprinzipien. Zweiter Bericht aus dem Forschungsprojekt „Modell-begründetes Lernen und Lehren. Multimediale Lernumgebungen als Gelegenheiten zum Nachdenken*. Freiburg: Institut für Erziehungswissenschaft.

Cortina, C. (2006). Psychologie der Umwelt. In A. Krapp & B. Weidenmann (Eds.), *Pädagogische Psychologie. Ein Lehrbuch* (S. 489-501). Weinheim: Beltz.

Eckert, A. (1998). *Kognition und Wissensdiagnose. Die Entwicklung und empirische Überprüfung des computerunterstützten wissensdiagnostischen Instrumentariums Netzwerk-Elaborierungs-Technik (NET).* Lengerich: Pabst Science Publishers.

Eels, W. C. (1930). Reliability of repeated grading of essay type examinations. *Journal of Educational Psychology*, 21, 48-52.

Eels, W. C. (1995). Die Zuverlässigkeit wiederholter Benotung von aufsatzähnlichen Prüfungsarbeiten. In K. Ingenkamp (Ed.), *Die Fragwürdigkeit der Zensurengebung: Texte und Unterrichtsberichte* (S. 167-172). Weinheim: Beltz.

Engel, F. & Hurrelmann, K. (1989). Psychosoziale Belastung im Jugendalter, Berlin: de Gruyter.

Engelkamp, J. (1994). Mentale Repräsentationen im Kontext verschiedener Aufgaben. In H. J. Kornadt, J. Grabowski & R. Mangold-Allwinn (Eds.), *Sprache und Kognition* (S. 37-54). Heidelberg: Spektrum.

Eigler, G. (1997). Textproduktion als konstruktiver Prozeß. In F. E. Weinert (Ed.), *Enzyklopädie der Psychologie - Pädagogische Psychologie. Band III: Psychologie des Unterrichts und der Schule* (S. 365-395). Göttingen: Hogrefe.

Etzioni, A. (Ed.) (1969). *Die Semi-Professions and their Organizations. Teachers, Nurses, Social workers.* New York: The Free Press.

Fend, H. (2008): Schule gestalten. Systemsteuerung, Schulentwicklung und Unterrichtsqualität. Wiesbaden: Verlag für Sozialwissenschaften.

Flick, U., Kardorff, E. & Steinke, I. (2010). Was ist qualitative Forschung? Einnleitung und Überblick, in U. Flick, E. Kardorff & I. Steinke (Eds.), Qualitative Forschung: Ein Handbuch (S. 13.29). Hamburg: Rowohlt.

Friedrich, H. F., & Mandl, H. (2006). Lernstrategien: Zur Strukturierung des Forschungsfeldes. In H. Mandl & H. F. Friedrich (Eds.), *Handbuch Lernstrategien* (S. 1-23). Göttingen: Hogrefe.

Gläser-Zikuda, M. (2007). Training selbstregulierten Lernens auf der Basis des Portfolio-Ansatzes. In M. Landmann & B. Schmitz (Eds.), *Selbstregulation erfolgreich fördern. Praxisnahe Trainingsprogramme für effektives Lernen* (S. 111-130). Stuttgart: Kohlhammer.

Gläser-Zikuda, M. (2008). Zum Ertrag Qualitativer Inhaltsanalyse in Pädagogik und Psychologie. In P. Mayring & M. Gläser-Zikuda (Eds), Die Praxis der Qualitativen Inhaltsanalyse. Weinheim: Beltz 2005, S. 186-296.

Gläser-Zikuda, M., Rohde, J. & Schlomske, N. (2010). Empirische Studien zum Lerntagebuch und Portfolio-Ansatz im Bildungskontext - ein Überblick. In M. Gläser-Zikuda (Ed.), *Lerntagebuch und Portfolio aus empirischer Sicht* (S. 3-34). Landau: Empirische Pädagogik.

Gläser-Zikuda, M., Seidel, T., Rohlfs, C., Gröschner, A. & Ziegelbauer, S. (2012). Mixed Methods in der empirischen Bildungsforschung – eine Einführung in die Thematik. In M. Gläser-Zikuda, T. Seidel, C. Rohlfs, A. Gröschner. & S. Ziegelbauer (Eds.), *Mixed Methods in der empirischen Bildungsforschung* (S. 7-13). Münster: Waxmann.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs**.** *Educational Evaluation and Policy Analysis*, *11,* 255–274.

Gruber, H. (1994). *Expertise. Modelle und empirische Untersuchungen.* (Bd. 34). Opladen: Westdeutscher Verlag.

Gruber, H. & Mandl, H. (1996). Das Entstehen von Expertise. In J. Hoffmann & W. Kintsch (Eds.), *Enzyklopädie der Psychologie, Themenbereich C, Theorie und Forschung, Serie II, Kognition,* (Bd. 7, S. 583-615). Göttingen: Hogrefe.

Hager, W. & Hasselhorn, M. (2008). Pädagogisch-psychologische Interventionsmaßnahmen. In W. Schneider & M. Hasselhorn (Eds.), *Handbuch der Pädagogischen Pychologie* (S. 339-347). Göttingen: Hogrefe.

Hagenauer, G. (2010). Pädagogisch-psychologische Interventionsmaßnahmen. In T. Hascher & B. Schmitz (Eds.), *Pädagogische Interventionsforschung. Theoretische Grundlagen und empirisches Handlungswissen* (S. 243-251). Weinheim: Juventa.

Hanke, U. (2006). *Externale Modellbildung als Hilfe bei der Informationsverarbeitung und beim Lernen.* Freiburg: Universitäts-Dissertation.

Hartmann, W. & Lehmann, R. (1989). Schüleraufsätze international untersucht. *Uni-hh Forschung* 23, 10-14.

Hascher, T. (2010). Unterschiedliche Interventionsimpulse: forschungs- versus praxisorientierte Ansätze. In T. Hascher & B. Schmitz (Eds.), *Pädagogische Interventionsforschung. Theoretische Grundlagen und empirisches Handlungswissen* (S. 269-279). Weinheim: Juventa.

Heiden, U. an der (1985). Kognitive Selbstreferenz. Band 1. In G. Pasternack (Ed.), *Erklären, Verstehen, Begründen* (S. 59-86). Bremen: Universität.

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität – Diagnose, Evaluation und Verbesserung des Unterrichts.* Seelze: Klett-Kallmeyer.

Helmke, A. & Schrader, F.-W. (2002). Jenseits von TIMSS: Messungen sprachlicher Kompetenzen, komplexe Längsschnittstudien und kulturvergleichende Analysen. Ergebnisse und Perspektiven ausgewählter Leistungsstudien. In F. E. Weinert (Ed.), *Leistungsmessungen in Schulen* (S. 237-250). Weinheim: Beltz.

Helmke, A. & Weinert, F. E. (1997). Bedingungsfaktoren schulischer Leistungen. In F. E. Weinert (Ed.), *Enzyklopädie der Psychologie, Band 3 (Psychologie der Schule und des Unterrichts)* (S. 71-176). Göttingen: Hogrefe.

Hoge, R. & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A Review of literature. *Review of Educational Research,* 59, 297–313.

Ifenthaler, D. (2006). *Diagnose lernabhängiger Veränderung mentaler Modelle. Veränderungsmessungen als Verfahren der empirischen Lehr-Lern-Forschung.* Freiburg: Universitäts-Dissertation.

Ifenthaler, D. (2008). Relational, structural, and semantic analysis of graphical representations and concept maps. *Educational Technology Research and Development*., *58* (1); S. 81-97.

Ifenthaler, D. (2010). Scope of graphical indices in educational diagnostics. In D. Ifenthaler, P. Pirnay-Dummer & N. M. Seel (Eds*.), Computer-Based Diagnostics and Systematic Analysis of Knowledge* (S. 213-234). New York: Springer.

Ifenthaler, D. & Pirnay-Dummer, P. (2009). *The overestimated construct validity in Knowledge Assessment Method.* Paper presented at the AERA, Devision C, Section 6: Learning and Instruction, San Diego, CA, 2008.

Ingenkamp, K. (1981). Die Messung und Bewertung des Lernerfolges in der Schule. In W. Twellmann (Ed.), *Handbuch Schule und Unterricht* (Bd. 1, S. 308-328). Düsseldorf. Schwann.

Ingenkamp, K. (1989). *Die Fragwürdigkeit der Zensurengebung*. Weinheim: Beltz.

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

Ingenkamp, K. & Lissmann, U. (2008). *Lehrbuch der pädagogischen Diagnostik. Studienausgabe*. Weinheim und Basel: Beltz.

Jäger, R. S. (2009). Diagnostische Aufgaben und Kompetenzen von Lehrkräften. In K.-H. Arnold, U. Sandfuchs & R. Wiechmann (Eds.), *Handbuch Unterricht* (S. 471-476). Bad Heilbrunn: Klinkhardt.

Johnson-Laird, P. N. (1983). *Mental models. Towards a cognitive science of language, inference, and consciousness.* Cambridge, UK: Cambridge University Press.

Klauer, K. J. (1978). *Handbuch der Pädagogischen Diagnostik (Studienausgabe),* Band 3. Düsseldorf: Schwann.

Klauer, K. J. (1982). *Handbuch der Pädagogischen Diagnostik (Studienausgabe),* Band 2. Düsseldorf: Schwann.

Klauer, K. J. (2002). Wie misst man Schulleistungen? In F. E. Weinert (Ed.), *Leistungsmessungen in Schulen* (S. 103-115). Weinheim: Beltz.

Koul, R., Clariana, R. B., & Salehi, R. (2005). Comparing several human and computer-based methods for scoring concept maps and essays. *Journal of Educational Computing Research, 32* (3), 261-273.

Krolak-Schwerdt, S., Böhmer, M. & Gräsel, C. (2012). Leistungsbeurteilungen von Schulkindern. Welche Rolle spielen Ziele und Expertise der Lehrkraft? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *44* (3), 111-122.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284.

Lenhard, W., Baier, H., Hoffmann, J. & Schneider, W. (2007). *Automatische Bewertung offener Antwortformate mittels Latenter Semantischer Analyse*. In W. Lenhard, E. Breitenbach, J. Schindelhauer-Deutscher, K. Zang & W. Henn (Eds.), (S. 155-165). *Diagnostica*, *53*, 155-165.

Leutner, D. (2010). Pädagogisch-psychologische Diagnostik. In D. H. Rost (Ed.), Handwörterbuch Pädagogische Psychologie (S. 624-635). Weinheim: PVU.

Lienert, G. A. & Raatz, U. (1994). *Testaufbau und Testanalyse*. Weinheim: Beltz.

Linn, R. L., Klein, S. P. & Hart, F. M. (1970). The nature and correlates of law school essay grades. (*Research Bulletin* 1970-4). Princeton: ETS. Kolumnentitel 27.

Mayring, P. (2008a). Neuere Entwicklungen in der qualitativen Forschung und der Qualitativen Inhaltsanalyse. In P. Mayring & M. Gläser-Zikuda (Eds.*), Die Praxis der Qualitativen Inhaltsanalyse* (S. 7-19). Weinheim: Beltz.

Mayring, P. (2008b). *Qualitative Inhaltsanalyse. Grundlagen und Techniken.* Weinheim: Beltz Verlag.

Mayring, P. (2012). Qualitative Inhaltsanalyse – ein Beispiel für Mixed Methods, In M. Gläser-Zikuda, T. Seidel, C. Rohlfs, A. Gröschner. & S. Ziegelbauer (Eds.), *Mixed Methods in der empirischen Bildungsforschung* (S. 27-36). Münster: Waxmann.

Mayring, P. & Gläser-Zikuda, M. (Eds.). (2008). *Die Praxis der Qualitativen Inhaltsanalyse.* Weinheim: Beltz.

McElvany, N., Schroeder, S., Hachfeld, A., Baumert, J., Richter, T., Schnotz, W. et al. (2009). Diagnostische Fähigkeiten von Lehrkräften bei der Einschätzung von Schülerleistungen und Aufgabenschwierigkeiten bei Lernmedien mit instruktionalen Bildern. *Zeitschrift für Pädagogische Psychologie, 23*, 223-235.

Meer, E. van. der (1996). Gesetzmäßigkeiten und Steuerungsmöglichkeiten des Wissenserwerbs. In *Enzyklopädie der Psychologie*, Bd. 2, S. 209-241. Göttingen: Hogrefe.

Mieg, H. (2005). Professionalisierung. In F. Rauner (Ed.), *Handbuch der Berufsbildungsforschung* (S. 342-348). Bielefeld: Bertelsmann.

Oevermann, U. (1996). Theoretische Skizze einer revidierten Theorie professionalisierten Handelns. In A. Combe & W. Helsper (Eds.), *Pädagogische Professionalität. Untersuchungen zum Typus pädagogischen Handelns* (S. 70-182). Frankfurt: Suhrkamp.

Piaget, J. (1976). *Die Äquilibration der kognitiven Strukturen*. Stuttgart: Klett.

Pirnay-Dummer, P. (2006). *Expertise und Modellbildung. MITOCAR.* Freiburg: Universitäts-Dissertation.

Pirnay-Dummer, P., 2007. *Model inspection trace of concepts and relations. A heuristic approach to language-oriented model assessment*. Paper presented at the AERA 2007, Division C, TICL SIG, April 2007, Chicago.

Pirnay-Dummer. P. (2010). Complete structure comparison. In D. Ifenthaler, P. Pirnay-Dummer & N. M. Seel (Eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge* (S. 235-258). New York: Springer.

Pirnay-Dummer, P. (2011). Comparison measures of T-MITOCAR, HIMATT, and AKOVIA. Retrieved March 30, 2011, from http://www.pirnaydummer.de/research/comparison_measures_2011-03-30.pdf

Pirnay-Dummer, P. (2012). *Die Sprache des Lernens. Theoretische Grundlagen, empirische Untersuchungen und Technologien.* Habitilationsschrift, Albert-Ludwigs-Universität, Freiburg.

Pirnay-Dummer, P., & Ifenthaler, D. (2010). Automated Knowledge Visualization and Assessment. In D. Ifenthaler, P. Pirnay-Dummer & N. M. Seel (Eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge* (S. 77-115). New York: Springer.

Pirnay-Dummer, P., Ifenthaler, D. & Spector, J. M. (2010). Highly integrated model assessment technology and tools. *Educational Technology Research and Development.,* 58 (1), 3-18.

Pirnay-Dummer, P. & Spector, J. M. (2008*). Language, association, and model rerepresentation. How features of language and human association can be utilized for automated knowledge assessment.* Paper presented to the AERA 2008, TICL SIG.

Pirnay-Dummer, P., & Walter, S. (2009). Bridging the world's knowledge to individual knowledge using latent semantic analysis and web ontologies to complement classical and new knowledge assessment Technologies. *Technology, Instruction, Cognition and Learning, 7* (1), 21-45.

Pintrich, P. R. (1988). A process-oriented view of student motivation and cognition. In J. Stark & L. Mets (Eds*.), Improving teaching and learning through research* (S. 65-79). San Franciso: Jossey-Bass.

Pohlmann, B. & Möller, J. (2007). Assimilations- und Kontrasteffekte bei der Bewertung von Texten. *Zeitschrift für Pädagogische Psychologie*, 21*, 297-303.*

Posner, M. (1988). Introduction: What is it to be an expert? In M. Chi, R. Glaser & M. Farr (Eds.), *The nature of expertise* (S. xxix–xxxvi). Hillsdale, NJ: Erlbaum.

Prenzel, M.; Seidel, T.; Lehrke, M.; Rimmele, R.; Duit, R.; Euler, M. et al. (2002). Lehr- Lernprozesse im Physikunterricht – eine Videostudie. In M. Prenzel; J. Doll (Eds.), *Bildungsqualität von Schule: Schulische und ausserschulische Bedingungen mathematischer, naturwissenschaftlicher und überfachlicher Kompetenzen* (S. 139–156)*. Zeitschrift für Pädagogik*, 45. Beiheft. Weinheim: Beltz.

Reinisch, H. (2009). „Lehrerprofessionalität" als theoretischer Term – Eine begriffssystematische Analyse. In O Zlatkin-Troitschanskaia, K. Beck, D. Sembill, R. Nickolaus & R. Mulder (Eds.), *Lehrerprofessionalität – Bedingungen, Genese, Wirkungen und ihre Messung* (S. 33-44). Weinheim: Beltz.

Rheinberg, F. (1978). Gefahren Pädagogischer Diagnostik. In K. J. Klauer (Ed.), *Handbuch der Pädagogischen Diagnostik* (Bd.1, S. 27-38). Düsseldorf: Schwann.

Rheinberg, F. (2002). Bezugsnormen und schulische Leistungsbeurteilung. In F. E. Weinert (Ed.), *Leistungsmessungen in Schulen* (S. 59-71). Weinheim: Beltz.

Rheinberg, F. (2006). Bezugsnorm-Orientierung. In D. H. Rost (Ed.), *Handwörterbuch Pädagogische Psychologie* (Bd. 3, S. 55-62). Weinheim: Beltz.

Rheinberg, F. (2008). Bezugsnormen und die Beurteilung von Lernleistung. In W. Schneider & M. Hasselhorn (Eds.), *Handbuch der Pädagogischen Psychologie* (S. 178-186). Göttingen: Hogrefe.

Rheinberg, F. (2009). Bezugsnormorientierung. In K.-H. Arnold, U. Sandfuchs & R. Wiechmann (Eds.), *Handbuch Unterricht* (S. 479-483). Bad Heilbrunn: Klinkhardt.

Rheinberg, F. & Fries, S. (2010). Bezugsnormorientierung. In D. H. Rost (Ed.), *Handwörterbuch Pädagogische Psychologie* (S. 61-68). Weinheim: Beltz.

Rumelhart, D. E. & Norman, D. A. (1978). Accretion, tuning, and restructuring: Three modes of learning. In J. Cotton & R. Klatzky (Eds.), Semantic factors in cognition. (S.37-53). Hillsdale, NJ: Erlbaum.

Schlomske, N., & Pirnay-Dummer, P. (2009). Model based assessment of learning dependent change within a two semester class. *Educational Technology Research and Development.,* 57 (6), 753-765.

Schmidt, M. & Otto, B. (2010). Direkte und indirekte Interventionen. In T. Hascher & B. Schmitz (Eds.), *Pädagogische Interventionsforschung. Theoretische Grundlagen und empirisches Handlungswissen* (S. 235-242). Weinheim: Juventa.

Schmitz, B. (2001). Self-Monitoring als transferfördernde Maßnahme bei einem Training zur Selbstregulation für Studierende. Eine prozessanalytische Untersuchung. *Zeitschrift für Pädagogische Psychologie*, 15, 179-195.

Schmitz, B., Landmann, M. & Perels, F. (2007). Das Selbstregulations-prozessmodell und theoretische Implikationen. In: M. Landmann & B. Schmitz (Eds.), *Selbstregulation erfolgreich fördern. Praxisnahe Trainingsprogramme für effektives Lernen* (S.312-327). Stuttgart: Kohlhammer.

Schmitz, B. & Schmidt, M. (2007). Einführung in die Selbstregulation. In M. Landmann & B. Schmitz (Eds.), *Selbstregulation erfolgreich fördern, Praxisnahe Trainingsprogramme für effektives Lernen (S. 9-19)*. Stuttgart: Kohlhammer.

Schmitz, B., & Wiese, B. .S. (2006). New perspectives fort he evaluation of training sessions in self-regulated learning: Time-series analysis of diary data. *Contemporary Educational Psychology, 31,* 64-96.

Schnotz, W. (1994). *Aufbau von Wissensstrukturen. Untersuchungen zur*

Kohärenzbildung beim Wissenserwerb mit Texten. Weinheim: Beltz Psychologie Verlags Union.

Schrader, F.-W. (2006). Diagnostische Kompetenz von Eltern und Lehrern. In D. H. Rost (Ed.), *Handwörterbuch Pädagogische Psychologie* (Vol. 3, S. 95-100). Weinheim: Beltz.

Schrader, F.-W. & Helmke, A. (2002). Alltägliche Leistungsbeurteilung durch Lehrer. In F. E. Weinert (Ed.), *Leistungsmessungen in Schulen* (S. 45-58). Weinheim: Beltz.

Schuler, H. (2010). Noten als Prädiktoren von Studien- und Berufserfolg. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 599-606). Weinheim: Beltz.

Schwarzer, C. (1982). *Einführung in die pädagogische Diagnostik.* München: Kösel.

Schwarzer, C. & Buchwald, P. (2006). Beratung in Familie, Schule und Beruf. In A. Krapp & B. Weidenmann (Eds.), *Pädagogische Psychologie* (S. 575-612). Weinheim: Beltz.

Shulman, L.S. (1986). *Those who understand. Knowledge growth in teaching.* In Educational Researcher 15, (2) S. 4-14.

Shulman, L. S. (2000). Teacher development: Roles of domain expertise and pedagogical knowledge. Journal of Applied Development Psychology, 21 (1), 129-135.

Seel, N. M. (1991). *Weltwissen und mentale Modelle.* Göttingen: Hogrefe.

Seel, N.M. (2003). *Psychologie des Lernens. Lehrbuch für Pädagogen und Psychologen*. München: Reinhardt.

Seel, N. M. (2010). Essentials for computer-based diagnostics of learning and cognition. In D. Ifenthaler, P. Pirnay-Dummer & N. M. Seel (Eds.), *Computer- Based Diagnostics and Systematic Analysis of Knowledge* (S. 3-14). New York: Springer.

Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher*, 15(2), 4–14.

Shulman, L. S. (2000). Teacher development: Roles of domain expertise and pedagogical knowledge. *Journal of Applied Developmental Psychology*, 21(1), 129–135.

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer/innen und das Konstrukt der diagnostischen Kompetenz. *Zeitschrift für Pädagogische Psychologie, 19 (1-2), 85-95.*

Stachowiak, H. (1973). *Allgemeine Modelltheorie*. Wien: Springer.

Straka, G. A. (2006). Lernstrategien in Modellen selbst gesteuerten Lernens. In H. Mandl & H. F. Friedrich (Eds.), *Handbuch Lernstrategien*. Göttingen: Hogrefe, S. 390-404.

Tashakkori, A. & Teddlie, C. (Eds.) (1998). *Mixed methodology: combining the qualitative and quantitative approaches.* Thousand Oaks, CA: Sage Publications, Inc

Teddlie, C.*, &* Tashakkori, A. *(*2010*).* Overview of contemporary issues in mixed methods research. In A. Tashakkori *&* C. Teddlie (Eds.), *Sage handbook of mixed methods in social & behavioral research (*2*nd ed., S.* 1-41*).* Thousand Oaks, CA*:* SAGE.

Tent, L. (2006). Zensuren. In D. H. Rost (Ed.), *Handwörterbuch Pädagogische Psychologie* (S. 873-880). Weinheim: Beltz.

Tent, L. & Birkel, P. (2010). Zensuren. In D. H. Rost (Ed.), *Handwörterbuch Pädagogische Psychologie* (S. 949-958). Weinheim: Beltz.

Trapmann, S., Hell, B., Weigand, S., & Schuler, H. (2007). Die Validität von Schulnoten zur Vorhersage des Studienerfolgs – eine Metaanalyse. *Zeitschrift für Pädagogische Psychologie*, 21, 11-27.

Trittel, M. (2010). Einzelfallanalysen und Studien mit kleinen Fallzahlen. In T. Hascher & B. Schfmitz (Eds.), *Pädagogische Interventionsforschung. Theoretische Grundlagen und empirisches Handlungswissen* (S. 280-286). Weinheim: Juventa.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327-352.

Weinert, F. E. (2002). Perspektiven der Schulleistungsmessung – Mehrperspektivistisch betrachtet. In F. E. Weinert (Ed.), *Leistungsmessungen in Schulen* (S. 353-365). Weinheim: Beltz.

Weinert, F. E., Schrader, F.-W. (1986). Diagnose des Lehrers als Diagnostiker. In H. Petillon, J. Wagner & B. Wolf (Eds.), *Schülergerechte Diagnose. Theoretische und empirische Beiträge zur Pädagogischen Diagnostik* (S. 11-29). Weinheim: Beltz.

Weinstein, C. E. & Mayer, R. E. (1986). The teaching of learning strategies. In M. Wittrock (Ed.), *Handbook of research on teaching* (S. 315-327). New York: Macmillian.

Weiss, R. (1995). Die Zuverlässigkeit der Ziffernbenotung bei Aufsätzen und Rechenarbeiten. In K. Ingenkamp (Ed.), *Die Fragwürdigkeit der Zensurengebung: Texte und Untersuchungsberichte* (S. 104- 116). Weinheim: Beltz.

Wessels, Michael G. (1994). Kognitive Psychologie. München: Reinhardt.

Wild, K. P. (2001). Die Optimierung von Videoanalysen durch zeitsynchrone Befragungsdaten aus dem Experience Sampling. In: S. v. Aufschnaiter; M. Welzel (Hrsg.), *Nutzung von Videodaten zur Untersuchung von Lehr-Lernprozessen: Aktuelle Methoden empirischer pädagogischer Forschung* (S. 61–74). Münster: Waxmann.

Wild, K.-P. & Rost, D. H. (1995). Klassengröße und Genauigkeit von Schülerbeurteilungen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie,* 27 (1), 78- 90.

Willett, J. B. (1989). Questions and answers in the measurement of change. In E. Z. Rothkopf (Ed.), *Review of Research in Education*, Volume 15. Washington, D.C.: American Education Research Association, 345-422.

Winter, F. (2008): Leistungsbewertung. Eine neue Lernkultur braucht einen anderen Umgang mit den Schülerleistungen. Baltmannsweiler: Schneider 2004.

Ziegenspeck, J. (1999). Handbuch Zensur und Zeugnis in der Schule. Historischer Rückblick, allgemeine Problematik, empirische Befunde und bildungspolitische Implikationen. Bad Heilbrunn: Klinkhardt.

Zimmerman, B. J. (2000). Attaining self-regulation. A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (S. 13-39). San Diego: Academic Press.

Zimmermann, B. J. (2005). The hidden dimension of personal competence. Self-regulated learning and practice. In A. J. Elliot & C. S. Dweck (Eds.), Handbook of competence and motivation (S. 509-526). New York: Guilford Press.

Abbildungsverzeichnis

Tabellenverzeichnis

Anhang

A Interviewleitfaden

Vorgehensweise und Bewertungskriterien

- Nach welchen Kriterien bewerten Sie textbasierte Schülerleistungen?
- Erklären Sie mal, wie Sie vorgegangen sind - von der Aufgabenentwicklung bis zur Bewertung.

Berücksichtigung der Gütekriterien

Objektivität

- Durchführungsobjektivität
    - Welche Bearbeitungszeit geben Sie Schülern, die in bestimmten Bereichen ein Defizit haben (beispielsweise Lese-Rechtschreibschwäche oder Sprachschwierigkeiten)?
    - Welche Hinweise geben Sie Ihren Schülern beim Austeilen der Klausuren? Wie reagieren Sie auf Nachfragen?
    - Welchen Eindruck haben Sie, ob Ihre Schüler beim Bearbeiten der Klausuren unter gleichen Rahmenbedingungen schreiben?
- Auswertungsobjektivität
    - Welche Bewertungskriterien kennen Ihre Schüler bevor die Klausur geschrieben wird?
- Interpretationsobjektivität
    - Erzählen Sie mal, wie ihre Bewertungsmaßstäbe zusammengekommen sind.
- **Reliabilität**
    - Haben Sie schon mal versucht mehrere Bewerter heranzuziehen - beispielsweise in „schwierigen“ Fällen? Wie sind Sie dabei vorgegangen?
- **Inhaltsvalidität**
    - Welche Inhalte erfassen Ihre Klausuren?

Kenntnis und Vermeidung der Fehlerquellen

Wie minimieren Sie Bewertungsfehler?

Reihungsfehler

- Welchen Eindruck haben Sie von einer Schülerleistung wenn Sie unmittelbar davor eine sehr gute/ sehr schlechte Klausur bewertet haben?

Milde/ Strenge

- Haben Sie den Eindruck, dass Ihre Kollegen Ihre Schüler strenger oder milder bewerten als Sie? Woran machen Sie das fest?

Haloeffekt

- Hat Ihrer Meinung nach das Schriftbild einen Einfluss auf Ihre Bewertung? Wenn ja, welchen?

Pygmalioneffekt

- Wenn ein Schüler Ihnen sympathisch/ unsympathisch ist, welchen Einfluss hat diese auf Ihre Bewertung?

Dimensionen diagnostischer Lehrerurteile

- Welche Aufgabenniveaus berücksichtigen Sie in Ihrer Klausur?
- An welchem Maßstab orientieren Sie sich bei der Bewertung der Schülerleistungen?

B Kodierleitfaden

Tabelle 99 Kodierleitfaden

Variable	Ausprägung	Definition	Ankerbeispiel	Kodierregeln
Durchführungs-objektivität	starke Ausprägung	Die Lehrkraft achtet darauf, dass die Schüler unter gleichen Rahmenbedingungen schreiben. Sie nennt Maßnahmen und konkrete Beispiele, um dies zu gewährleisten.	„Ich sage meinen Schülern vor der Klausur, welche Hilfsmittel sie verwenden dürfen. Dies tue ich gewöhnlich schon bei der Bekanntgabe des Klausurtermins und nicht erst zur Klausur.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „niedrige Ausprägung“ enthalten sein.
	mittlere Ausprägung	Die Lehrkraft achtet darauf, dass die Schüler unter gleichen Rahmenbedingungen schreiben. Sie nennt keine Maßnahmen und keine konkreten Beispiele, um dies zu gewährleisten.	„Ja, die schreiben unter den gleichen Bedingungen.“	Sobald über diesen Aspekt gesprochen wird, ist er nicht mehr stark oder niedrig ausgeprägt.
	niedrige Ausprägung	Die Lehrkraft achtet nicht darauf, dass die Schüler unter gleichen Rahmenbedingungen schreiben.	„Darauf achte ich nicht.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „starke Ausprägung“ enthalten sein.
	nicht erschließbar			

Variable	Ausprägung	Definition	Ankerbeispiel	Kodierregeln
Durchführungs-objektivität	starke Ausprägung	Die Lehrkraft greift während der Durchführung der Klausur überhaupt nicht ein: - reagiert nicht auf Nachfragen - gibt keine Hinweise (weder inhaltlicher noch formeller Art)	„Die Aufgabe muss eindeutig formuliert sein. Ich reagiere nicht auf Nachfragen“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „niedrige Ausprägung“ enthalten sein.
	mittlere Ausprägung	Die Lehrkraft gibt bei der Durchführung der Klausur ausschließlich Hilfestellung zu formellen Aspekten der Klausur.	„Wenn es Fragen hinsichtlich meiner Schrift sind, gebe ich Hinweise“	Sobald über diesen Aspekt gesprochen wird, ist er nicht mehr stark oder niedrig ausgeprägt.
	niedrige Ausprägung	Die Lehrkraft gibt bei der Durchführung der Klausur Hilfestellung zu inhaltlichen - und formellen Aspekten der Klausur.	„Ich erinnere an die Gruppenarbeit, die wir im Unterricht gemacht haben.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „starke Ausprägung“ enthalten sein.
	nicht erschließbar			

Variable	Ausprägung	Definition	Ankerbeispiel	Kodierregeln
Durchführungs-objektivität	starke Ausprägung	Die Lehrkraft kennt und präzisiert die Regelungen bei Schülern mit Lese-/ Rechtschreibschwäche oder Sprachschwierigkeiten und wendet diese an. Dabei kann sie die präzise Regelung nennen.	„Ja, die gibt es, diese Sonderregelung, dass diese Schüler 15 Minuten länger schreiben dürfen.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „niedrige Ausprägung“ enthalten sein.
	mittlere Ausprägung	Die Lehrkraft kennt die Regelungen (ohne Präzision) bei Schülern mit Lese-Rechtschreibschwäche oder Sprachschwierigkeiten und wendet diese an.	„Ja, denen gebe ich extra Zeit.“	Sobald über diesen Aspekt gesprochen wird, ist er nicht mehr stark oder niedrig ausgeprägt.
	niedrige Ausprägung	Die Lehrkraft ist sich solch einer Regelung nicht bewusst und wendet sie nicht an.	„Bei mir bekommen alle die gleiche Zeit- ohne Ausnahmeregeln.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „starke Ausprägung“ enthalten sein.
	nicht erschließbar			

Variable	Ausprägung	Definition	Ankerbeispiel	Kodierregeln
Auswertungs-objektivität	starke Ausprägung	Die Schüler kennen die Bewertungskriterien, kennen den Notenschlüssel und sehen die Punkteverteilung in der Klausur.	„Meine Schüler kennen meinen Bewertungsbogen, da ich ihnen diesen im Vorfeld gezeigt habe.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „niedrige Ausprägung“ enthalten sein.
	mittlere Ausprägung	Die Schüler sehen zwar die Punkte, kennen aber den Notenschlüssel oder die Bewertungskriterien nicht.	„Die Punkte stehen an den Aufgabenstellungen dran.“	Sobald über diesen Aspekt gesprochen wird, ist er nicht mehr stark oder niedrig ausgeprägt.
	niedrige Ausprägung	Die Schüler kennen die Kriterien nicht, sehen die Punkte nicht und kennen auch den Notenschlüssel nicht.	„Die Schüler erfahren hinterher - bei der Klausurbesprechung - wie ich die Punkte verteilt habe.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „starke Ausprägung“ enthalten sein.
	nicht erschließbar			

Variable	Ausprägung	Definition	Ankerbeispiel	Kodierregeln
Interpretations-objektivität	starke Ausprägung	Verschiedene Lehrkräfte tauschen sich regelmäßig über ihre Bewertungskriterien aus. Es erfolgt ein konkretes Beispiel im Interview.	„Ja, wir tauschen uns darüber aus. Ich habe anderen Lehrkräften schon häufiger mal ein paar Klausuren gegeben um diese mit ihren Kriterien zu bewerten und herauszufinden wie ihr Gesamteindruck bezüglich der Leistung ist.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „niedrige Ausprägung“ enthalten sein.
	mittlere Ausprägung	Verschiedene Lehrkräfte tauschen sich regelmäßig über ihre Bewertungskriterien aus. Es erfolgt kein konkretes Beispiel im Interview.	„Wir haben uns mal darüber ausgetauscht. Aber eigentlich eher nicht.“	Sobald über diesen Aspekt gesprochen wird, ist er nicht mehr stark oder niedrig ausgeprägt.
	niedrige Ausprägung	Die Lehrkräfte tauschen sich nicht über ihre Bewertungskriterien aus.	„Die Zeit ist hierfür nicht gegeben.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „starke Ausprägung“ enthalten sein.
	nicht erschließbar			

Variable	Ausprägung	Definition	Ankerbeispiel	Kodierregeln
Interpretations-objektivität (siehe „Milde-Strenge“ Effekt im Interviewleitfaden)	starke Ausprägung	Die Lehrkraft ist sich über die „Milde/Strenge-Bewertung“ ihrer Kollegen bewusst und sie wendet konkrete Maßnahmen an und nennt sie.	„Nein, denn ich tausche mich mit meinen Kollegen über einzelne Noten aus und merke dabei, dass sie dieselbe Note oder eine ähnliche Note wie ich vergeben hätten.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „niedrige Ausprägung“ enthalten sein.
	mittlere Ausprägung	Die Lehrkraft ist sich über die „Milde/Strenge-Bewertung“ ihrer Kollegen bewusst. Sie nennt keine Maßnahmen.	„Nein wir bewerten eigentlich ähnlich, meine ich.“	Sobald über diesen Aspekt gesprochen wird, ist er nicht mehr stark oder niedrig ausgeprägt.
	niedrige Ausprägung	Die Lehrkraft ist sich über die „Milde/Strenge-Bewertung“ ihrer Kollegen nicht bewusst.	„Das weiß ich nicht, ob meine Kollegen strenger oder milder bewerten.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „starke Ausprägung“ enthalten sein.
	nicht erschließbar			

Variable	Ausprägung	Definition	Ankerbeispiel	Kodierregeln
Reliabilität	starke Ausprägung	Die Lehrkraft zieht häufig in spezifischen Fällen weitere Bewerter heran und hat dies bereits öfters getan.	„Ja das habe ich schon getan, vor allem wenn Schüler ihr Ergebnis auf Grundlage meiner Kriterien nicht nachvollziehen konnten.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „niedrige Ausprägung“ enthalten sein.
	mittlere Ausprägung	Die Lehrkraft zieht gelegentlich - in spezifischen Fällen - weitere Bewerter heran.	„Wir haben uns mal darüber ausgetauscht. Aber eigentlich eher nicht.“	Sobald über diesen Aspekt gesprochen wird, ist er nicht mehr stark oder niedrig ausgeprägt.
	niedrige Ausprägung	Die Lehrkraft zieht in spezifischen Fällen keine weiteren Bewerter heran.	„Dafür reicht die Zeit nicht.“	Mindestens 1 Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „starke Ausprägung“ enthalten sein.
	nicht erschließbar			

Variable	Ausprägung	Definition	Ankerbeispiel	Kodierregeln
Inhalts-validität	starke Ausprägung	Die Klausurinhalte beziehen sich auch tatsächlich auf das, was im Unterricht behandelt wurde. Der Lehrer kann die Maßnahme zur Validierung seines Tests erklären.	„Ich ziehe die konkret verwendeten Unterrichtsmaterialen der letzten Unterrichtseinheit heran. Gelegentlich gebe ich die Klausur - bevor sie geschrieben wird - auch einem Kollegen und hole mir so eine weitere Meinung ein.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „niedrige Ausprägung“ enthalten sein.
	mittlere Ausprägung	Die Klausurinhalte beziehen sich auch tatsächlich auf das, was im Unterricht behandelt wurde. Der Lehrer erklärt die Maßnahme zur Validierung seines Tests nicht.	„Die Inhalte, die im Unterricht thematisiert wurden.“	Sobald über diesen Aspekt gesprochen wird, ist er nicht mehr stark oder niedrig ausgeprägt.
	niedrige Ausprägung	Die Klausurinhalte beziehen sich nicht auf das, was im Unterricht behandelt wurde.	„Ich überlege nicht weiter was wir im Unterricht behandelt hatten, sondern konstruiere die Klausur auf Grundlage meines Bauchgefühls.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „starke Ausprägung“ enthalten sein.
	nicht erschließbar			

Variable	Ausprägung	Definition	Ankerbeispiel	Kodierregeln
Bewertungsfehler minimieren (Allgemein)	starke Ausprägung	Der Lehrkraft sind (allgemeine) Bewertungsfehler bewusst und sie wendet konkrete Maßnahmen an, um diese zu minimieren und nennt sie.	„Ich versuche diese zu umgehen, indem ich klare Kriterien im Vorfeld formuliere an denen ich mich orientiere.	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „niedrige Ausprägung“ enthalten sein.
	mittlere Ausprägung	Der Lehrkraft sind (allgemeine) Bewertungsfehler bewusst. Sie kennt keine Maßnahmen.	„Ja, das kann schon passieren. Ich versuche das allerdings zu vermeiden.“	Sobald über diesen Aspekt gesprochen wird, ist er nicht mehr stark oder niedrig ausgeprägt.
	niedrige Ausprägung	Der Lehrkraft sind (allgemeine) Bewertungsfehler nicht bewusst	„Bisher ist mir noch nicht aufgefallen, dass mir Fehler in der Bewertung unterlaufen wären.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „starke Ausprägung“ enthalten sein.
	nicht erschließbar			

Variable	Ausprägung	Definition	Ankerbeispiel	Kodierregeln
Reihungsfehler	starke Ausprägung	Der Lehrkraft ist der Reihungsfehler bewusst und sie wendet konkrete Maßnahmen an und nennt sie.	„Ich weiß, dass das in der Bewertung passieren kann und versuche dies bewusst zu umgehen indem ich die Klausuren in unterschiedlicher Reihenfolge korrigiere.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „niedrige Ausprägung“ enthalten sein.
	mittlere Ausprägung	Der Lehrkraft ist der Reihungsfehler bewusst. Sie nennt keine Maßnahmen.	„Das kann durchaus passieren.“	Sobald über diesen Aspekt gesprochen wird, ist er nicht mehr stark oder niedrig ausgeprägt.
	niedrige Ausprägung	Der Lehrkraft ist der Reihungsfehler nicht bewusst.	„Den Fehler gibt es nicht.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „starke Ausprägung“ enthalten sein.
	nicht erschließbar			

Variable	Ausprägung	Definition	Ankerbeispiel	Kodierregeln
Haloeffekt	starke Ausprägung	Der Lehrkraft ist der Haloeffekt bewusst und sie wendet konkrete Maßnahmen an und nennt sie.	„Ich bemühe mich, bewusst nicht auf das Schriftbild zu achten. Da ich mir dieses Fehlers bewusst bin, lese ich unleserliche Klausuren bewusst noch einmal durch, um mich selber zu überprüfen, ob ich keinen Punkt übersehen habe.	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „niedrige Ausprägung“ enthalten sein.
	mittlere Ausprägung	Der Lehrkraft ist der Haloeffekt bewusst. Sie nennt keine Maßnahmen.	„Ja, den Fehler kenne ich.“	Sobald über diesen Aspekt gesprochen wird, ist es nicht mehr stark oder niedrig ausgeprägt.
	niedrige Ausprägung	Der Lehrkraft ist der Haloeffekt nicht bewusst.	„Ich würde vermuten, dass es so einen Fehler nicht gibt.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „starke Ausprägung“ enthalten sein.
	nicht erschließbar			

Variable	Ausprägung	Definition	Ankerbeispiel	Kodierregeln
Pygmalioneffekt	starke Ausprägung	Der Lehrkraft ist der Pygmalioneffekt bewusst und sie wendet konkrete Maßnahmen an und nennt sie.	„Ich bemühe mich das zu vermeiden, dass das keinen Einfluss hat. Hierbei helfen mir meine Kriterien, die genau festlegen, was eine sehr gute Leistung ausmacht.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „niedrige Ausprägung“ enthalten sein.
	mittlere Ausprägung	Der Lehrkraft ist der Pygmalioneffekt bewusst. Sie nennt keine Maßnahmen.	„Ja, so einen Fall hatte ich schon mal, dass ein Schüler im Unterricht sehr gut mitgemacht hat und in der Klausur schlecht war.“	Sobald über diesen Aspekt gesprochen wird, ist er nicht mehr stark oder niedrig ausgeprägt.
	niedrige Ausprägung	Der Lehrkraft ist der Pygmalioneffekt nicht bewusst.	„Das passiert bei mir eigentlich nicht.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „starke Ausprägung“ enthalten sein.
	nicht erschließbar			

Variable	Ausprägung	Definition	Ankerbeispiel	Kodierregeln
Aufgabenniveaus	starke Ausprägung	Die Lehrkraft berücksichtigt bewusst unterschiedliche Aufgabenniveaus und nennt dabei ein konkretes Beispiel.	„Ja, in der ersten Aufgabe geht es meistens um einfache Reproduktion und in den darauffolgenden Aufgaben geht es um Verständnis und Transfer.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „niedrige Ausprägung“ enthalten sein.
	mittlere Ausprägung	Die Lehrkraft berücksichtigt bewusst unterschiedliche Aufgabenniveaus. Sie nennt dabei kein konkretes Beispiel.	„Ja, ich berücksichtige unterschiedliche Schwierigkeitsgrade.“	Sobald über diesen Aspekt gesprochen wird, ist er nicht mehr stark oder niedrig ausgeprägt.
	niedrige Ausprägung	Die Lehrkraft berücksichtigt nicht bewusst unterschiedlichen Aufgabenniveaus. Oder die Lehrkraft berücksichtigt bewusst keine unterschiedlichen Aufgabenniveaus.	„Darüber mache ich mir eigentlich keine Gedanken.“ „Ich weiß, es wäre sinnvoll, habe mich allerdings absichtlich dagegen entschieden.“	Mindestens ein Aspekt muss erfüllt sein und es darf dabei kein Aspekt „mittlere Ausprägung“ oder „starke Ausprägung“ enthalten sein.
	nicht erschließbar			

Variable	Definition	Ankerbeispiel	Kodierregeln
Kriteriale Bezugsnorm	Der Bewertungskriterienkatalog wird als Bezugsnorm herangezogen.	„ich bewerte auf Grundlage meiner Kriterien, die ich festgelegt habe“	Es wird primär die kriteriale Bezugsnorm angelegt.
Soziale Bezugsnorm	Der Klassenmaßstab wird als Referenz verwendet.	„Ich lege mir erst mal alle Klausuren in Stapel zurecht, sodass die sehr guten und die sehr schlechten Klausuren zusammen sind“	Es wird primär die soziale Bezugsnorm angelegt.
Individuelle Bezugsnorm	Die individuelle Bezugsnorm wird als Referenz verwendet.	„Das würde ich individuell entscheiden.“	Es wird primär die individuelle Bezugsnorm angelegt.
	nicht erschließbar		

C Detailliertere Ergebnisdarstellung der Hochschulstudien

Ergebnisdarstellung der ersten Hochschulstudie

Tabelle 100 Regressionsanalyse beider Bewerter hinsichtlich des selbstregulierten Lernens

			Musterlösung			Lehrtext		
Kriterien	**Kennwerte**	**df**	**Korr. R²**	ρ	**F**	**Korr. R²**	ρ	**F**
Inhalt	Struktur	4	-0.10	0.96	0.15	-0.01	0.45	0.94
Inhalt	Semantik	3	0.11	0.07	2.54	0.05	0.20	1.63

Tabelle 101 Korrelationskoeffizienten beider Bewerter hinsichtlich des selbstregulierten Lernens

	Kennwerte	**Musterlösung**	**Lehrtext**
Struktur	Surface Matching	0.01	0.13
	Graphical Matching	-0.10	0.12
	Structural Matching	0.02	-0.07
	Gamma Matching	0.20	0.19
Semantik	Concept Matching	0.40 (*)	0.09
	Propositional Matching	0.07	-0.27 (*)
	Balanced Semantic Matching	0.07	-0.27 (*)

Tabelle 102 Regressionsanalyse beider Bewerter hinsichtlich der Metakognition

			ML			GMM		
Kriterien	**Kennwerte**	**df**	**Korr. R²**	ρ	**F**	**Korr. R²**	ρ	**F**
Inhalt	Struktur	4	-0.08	0.88	0.30	-0.07	0.82	0.38
Inhalt	Semantik	3	-0.01	0.48	0.85	-0.03	0.59	0.65

ML= Musterslösung; GMM= Gesamtmodelll mit Stichpunkten

Tabelle 103 Regressionsanalyse beider Bewerter hinsichtlich der Metakognition

			GMO			LT		
Kriterien	**Kennwerte**	**df**	**Korr. R²**	ρ	**F**	**Korr. R²**	ρ	**F**
Inhalt	Struktur	4	-0.08	0.83	0.63	-0.08	0.82	0.39
Inhalt	Semantik	3	-0.04	0.65	0.56	-0.04	0.63	0.58

GMO= Gesamtmodelll ohne Stichpunkten; LT= Lehrtext

Tabelle 104 Korrelationskoeffizienten beider Bewerter hinsichtlich der Metakognition

	Kennwerte	ML	GMM	GMO	LT
Struktur	Surface Matching	0.03	0.15	0.13	0.15
	Graphical Matching	0.10	0.10	0.10	0.14
	Structural Matching	0.04	0.03	0.03	0.04
	Gamma Matching	-0.04	0.13	0.02	0.03
Semantik	Concept Matching	- 0.11	0.08	0.11	0.13
	Propositional Matching	0.06	-0.04	0.09	0.11
	Balanced Semantic Matching	0.04	-0.01	0.11	0.13

ML= Musterslösung; GMM= Gesamtmodelll mit Stichpunkten; GMO= Gesamtmodell ohne Stichpunkte; LT= Lehrtext

Ergebnisdarstellung der zweiten Hochschulstudie

Tabelle 105 Regressionsanalyse beider Bewerter hinsichtlich des selbstregulierten Lernens

			Musterlösung			GMO			Lehrtext		
Kriterien	**Kenn-werte**	**df**	**Korr. R²**	ρ	**F**	**Korr. R²**	ρ	**F**	**Korr. R²**	ρ	**F**
Inhalt	Struktur	4	0.01	-0.29	1.27	-0.03	0.85	0.35	0.01	-0.03	-0.03
Inhalt	Semantik	3	-0.02	0.82	0.31	0.04	0.09	2.56	0.05	0.06	2.59

GMO= Gesamtmodelll ohne Stichpunkten

Tabelle 106 Korrelationskoeffizienten beider Bewerter hinsichtlich des selbstregulierten Lernens

	Kennwerte	ML	GMO	LT
Struktur	Surface Matching	0.14	0.04	0.03
	Graphical Matching	- 0.01	0.002	0.02
	Structural Matching	0.11	0.06	0.06
	Gamma Matching	0.29 (*)	- 0.08	- 0.07
Semantik	Concept Matcinig	0.03	0.21 (*)	0.28 (*)
	Propositional Matching	- 0.03	0.08	0.15
	Balanced Semantic Matching	- 0.02	0.03	0.15

ML= Musterslösung; GMO= Gesamtmodelll mit Stichpunkten; LT= Lehrttext

Tabelle 107 Regressionsanalyse beider Bewerter hinsichtlich der Lernstrategien

			ML			GMO		
Kriterien	**Kennwerte**	**df**	**Korr. R²**	ρ	**F**	**Korr. R²**	ρ	**F**
Inhalt	Struktur	4	0.06	0.11	1.95	0.07	0.09	2.08
Inhalt	Semantik	3	0.08	0.07	2.52	0.06	0.09	2.25

ML= Musterslösung; GMO= Gesamtmodelll ohne Stichpunkten

Tabelle 108 Korrelationskoeffizienten beider Bewerter hinsichtlich der Lernstrategien

	Kennwerte	ML	GMO
Struktur	Surface Matching	0.06	- 0.12
	Graphical Matching	- 0.08	- 0.03
	Structural Matching	0.09	- 0.03
	Gamma Matching	0.28 (*)	0.23 (*)
Semantik	Concept Matching	0.19	0.28 (*)
	Propositional Matching	0.23	0.30 (*)
	Balanced Semantic Matching	0.21	0.28 (*)

ML= Musterslösung; GMO= Gesamtmodelll ohne Stichpunkten

Detailliertere Ergebnisdarstellung der Schulstudien

Ergebnisdarstellung im Unterrichtsfach Biologie

Tabelle 109 Regressionsanalyse der ersten Lehrkraft hinsichtlich der Musterlösung

			1 MP			2 MP		
Kriterien	**Kennwerte**	**df**	**Korr. R²**	ρ	**F**	**Korr. R²**	ρ	**F**
Inhalt	Struktur	4	0.16	0.08	2.35	0.24	0.15	1.85
Inhalt	Semantik	3	-0.03	0.57	0.70	-0.09	0.90	0.20
Layout	Struktur	4	0.06	0.25	1.55	0.06	0.24	1.48
Layout	Semantik	3	-0.02	0.49	0.83	0.14	0.08	2.53

MP= Messzeitpunkt

Tabelle 110 Regressionsanalyse der zweiten Lehrkraft hinsichtlich der Musterlösung

			1 MP			2 MP		
Kriterien	**Kennwerte**	**df**	**Korr. R²**	ρ	**F**	**Korr. R²**	ρ	**F**
Inhalt	Struktur	4	0.13	0.12	2.02	0.15	0.10	2.19
Inhalt	Semantik	3	-0.12	0.95	0.12	-0.08	0.83	0.29
Layout	Struktur	4	0.05	0.28	1.37	0.13	0.12	2.08
Layout	Semantik	3	0.09	0.50	0.82	0.10	0.45	0.91

MP= Messzeitpunkt

Tabelle 111 Regressionsanalyse der ersten Lehrkraft hinsichtlich des Gruppenmodells

			1 MP			2 MP		
Bewertungskriterien	**Kennwerte**	**df**	**Korr. R²**	ρ	**F**	**Korr. R²**	ρ	**F**
Inhaltlich orientiert	Struktur	4	-0.08	0.74	0.49	0.31	0.48	0.91
Inhaltlich orientiert	Semantik	3	-0.03	0.56	0.70	-0.09	0.90	0.20
Layout orientiert	Struktur	4	0.31	0.06	2.64	0.29	0.08	2.42
Layout orientiert	Semantik	3	-0.08	0.83	0.29	0.11	0.12	2.14

MP= Messzeitpunkt

Tabelle 112 Regressionsanalyse der zweiten Lehrkraft hinsichtlich des Gruppenmodells

			1 MP			2 MP		
Bewertungskriterien	**Kennwerte**	**df**	**Korr. R²**	ρ	**F**	**Korr. R²**	ρ	**F**
Inhaltlich orientiert	Struktur	4	0.03	0.96	0.16	-0.10	0.84	0.36
Inhaltlich orientiert	Semantik	3	-0.10	0.94	0.14	-0.00	0.42	0.98
Layout orientiert	Struktur	4	0.00	0.42	1.01	0.04	0.31	1.28
Layout orientiert	Semantik	3	-0.02	0.52	0.78	0.01	0.96	0.11

MP= Messzeitpunkt

Tabelle 113 Korrelationskoeffizienten der ersten Lehrkraft (inhaltlich orientierte Bewertung)

	Kennwerte	**1 MP**		**2 MP**	
		Lehrerlösung	Gruppenmodell	Lehrerlösung	Gruppenmodell
Struktur	Surface Matching	- 0.43 (*)	- 0.05	- 0.33 (*)	- 0.04
	Graphical Matching	- 0.17	- 0.23	- 0.22	- 0.22
	Structural Matching	- 0.13	- 0.06	- 0.12	0.05
	Gamma Matching	0.13	0.03	0.25	0.06
Semantik	Concept Matching	0.04	- 0.09	- 0.13	- 0.17
	Propositional Matching	0.09	- 0.09	0.04	- 0.04
	Balanced Semantic Matching	0.12	- 0.01	0.04	0.08

MP= Messzeitpunkt

Tabelle 114 Korrelationskoeffizienten der zweiten Lehrkraft (inhaltlich orientierte Bewertung)

	Kennwerte	**1 MP**		**2 MP**	
		Lehrerlösung	Gruppenmodell	Lehrerlösung	Gruppenmodell
Struktur	Surface Matching	- 0.14	- 0.05	- 0.15	0.04
	Graphial Matching	- 0.22	- 0.04	- 0.28	- 0.02
	Structural Matching	0.09	0.01	0.10	0.10
	Gamma Matching	0.34 (*)	- 0.02	0.32 (*)	0.10
Semantik	Concept Matching	0.09	0.12	0.06	0.22
	Propositional Matching	0.01	0.07	0.10	0.22
	Balanced Semantic Matching	0.02	0.00	0.09	0.20

MP= Messzeitpunkt

Tabelle 115 Korrelationskoeffizienten der ersten Lehrkraft (strukturell orientierte Bewertung)

	Kennwerte	**1 MP**		**2 MP**	
		Lehrerlösung	Gruppenmodell	Lehrerlösung	Gruppenmodell
Struktur	Surface Matching	- 0.43 (*)	0.12	- 0.34 (*)	- 0.20
	Graphical Matching	- 0.13	- 0.50 (*)	0.03	- 0.54 (*)
	Structural Matching	- 0.32 (*)	0.02	- 0.28	- 0.08
	Gamma Matching	0.10	- 0.02	- 0.20	- 0.06
Semantik	Concept Matching	0.24	- 0.09	0.21	0.04
	Propositional Matching	0.19	0.00	0.43 (*)	0.18
	Balanced Semantic Matching	0.18	- 0.09	0.38 (*)	0.07

MP= Messzeitpunkt

Tabelle 116 Korrelationskoeffizienten der zweiten Lehrkraft (strukturell orientierte Bewertung)

	Kennwerte	**1 MP**		**2 MP**	
		Musterlösung	Gruppenmodell	Musterlösung	Gruppenmodell
Struktur	Surface Matching	- 0.35 (*)	0.05	- 0.44 (*)	- 0.03
	Graphiccal Matching	- 0.28	- 0.14	- 0.35 (*)	- 0.29
	Structural Matching	- 0.39 (*)	- 0.19	- 0.35 (*)	0.15
	Gamma Matching	- 0.20	0.15	- 0.08	0.12
Semantik	Concept Matching	- 0.24	- 0.04	- 0.23	- 0.02
	Propositional Matching	- 0.10	0.17	- 0.15	- 0.01
	Balanced Semantic Matching	- 0.05	0.29	- 0.10	0.03

MP= Messzeitpunkt

Ergebnisdarstellung im Unterrichtsfach Deutsch

Tabelle 117 Regressionsanalyse des zweiten Messzeitpunktes (Musterlösung & Gesamtmodell)

Kriterien	Kennwerte	df	Korr. R²	ρ	F
Inhalt	Struktur	4	0.13	0.23	1.61
Inhalt	Semantik	3	-0.13	0.77	0.38
Sprache	Struktur	4	-0.24	0.92	2.23
Sprache	Semantik	3	0.15	0.17	1.97
Form	Struktur	4	-0.04	0.84	0.53
Form	Semantik	3	-0.09	0.65	0.57

Tabelle 118 Korrelationskoeffizienten (Musterlösung & Gesamtmodell)

		2 MP		
	Kennwerte	Inhalt	Sprache	Form
Struktur	Surface Matching	0.20	0.35	- 0.26
	Graphical Matching	0.03	0.25	0.31
	Structural Matching	-0.09	- 0.22	- 0.47 (*)
	Gamma Matching	0.57 (*)	0.03	0.08
Semantik	Concept Matching	0.15	0.13	- 0.02
	Propositional Matching	0.08	0.57 (*)	- 0.26
	Balanced Semantic Matching	0.13	0.58 (*)	- 0.29

MP= Messzeitpunkt

Ergebnisdarstellung im Unterrichtsfach Religion

Tabelle 119 Regressionsanalyse der ersten Lehrkraft (Gesamtbewertung)

			1 MP			**2 MP**		
	Kennwerte	**df**	**Korr. R²**	ρ	**F**	**Korr. R²**	ρ	**F**
Musterlösung	Struktur	4	-0.03	0.49	0.91	-0.21	0.82	0.38
Musterlösung	Semantik	3	-0.09	0.62	0.62	-0.09	0.62	0.61
Gesamtmodell	Struktur	4	0.18	0.71	0.55	-0.27	0.89	0.27
Gesamtmodell	Semantik	3	-0.21	0.90	0.19	0.26	0.11	2.60

MP= Messzeitpunkt

Tabelle 120 Korrelationskoeffizienten der ersten Lehrkraft (Gesamtbewertung)

	Kennwerte	**1 MP**		**2 MP**	
		Lehrerlösung	Gruppenmodell	Lehrerlösung	Gruppenmodell
Struktur	Surface Matching	0.05	0.16	- 0.05	0.30
	Graphical Matching	- 0.03	0.18	0.05	0.13
	Structural Matching	0.51 (*)	0.37	0.18	0.18
	Gamma Matching	0.44	0.15	0.32	0.22
Semantik	Concept Matching	0.05	- 0.002	- 0.14	- 0.34
	Propositional Matching	0.09	0.06	0.26	0.17
	Balanced Semantic Matching	- 0.02	0.16	0.32	0.45 (*)

MP= Messzeitpunkt

Tabelle 121 Regressionsanalyse der zweiten Lehrkraft (Gesamtbewertung)

			1 MP			**2 MP**		
	Kennwerte	**df**	**Korr. R²**	ρ	**F**	**Korr. R²**	ρ	**F**
Musterlösung	Struktur	4	0.50	0.06	3.67	0.45	0.08	3.27
Musterlösung	Semantik	3	0.20	0.20	1.94	0.07	0.35	1.26
Gesamtmodell	Struktur	4	0.55	0.04	4.41	0.38	0.12	2.68
Gesamtmodell	Semantik	3	0.10	0.31	1.41	-0.04	0.50	0.86

MP= Messzeitpunkt

Tabelle 122 Korrelationskoeffizienten der zweiten Lehrkraft (Gesamtbewertung)

	Kennwerte	**1 MP**		**2 MP**	
		Musterlösung	Gruppenmodell	Musterlösung	Gruppenmodell
Struktur	Surface Matching	0.42	0.78 (*)	0.73 (*)	0.73 (*)
	Graphical Matching	0.74 (*)	0.74 (*)	0.78 (*)	0.78 (*)
	Structural Matching	0.17	0.66 (*)	0.13	0.64 (*)
	Gamma Matching	0.17	0.60 (*)	0.20	0.73 (*)
Semantik	Concept Matching	0.12	0.50	0.24	0.47
	Propositional Matching	0.43	0.59 (*)	0.44	0.33
	Balanced Semantic Matching	0.54 (*)	0.48	0.45	0.11

Ergebnisdarstellung im Unterrichtsfach Kunst

Tabelle 123 Regressionsanalyse hinsichtlich der Musterlösung (Gesamtbewertung)

			1 MP			**2 MP**		
	Kennwerte	**df**	**Korr. R²**	ρ	**F**	**Korr. R²**	ρ	**F**
Musterlösung	Struktur	4	0.08	0.18	1.62	-0.07	0.75	0.48
Musterlösung	Semantik	3	0.01	0.36	1.12	-0.02	0.49	0.84
Gesamtmodell	Struktur	4	-0.002	0.43	0.98	0.02	0.34	1.18
Gesamtmodell	Semantik	3	0.01	0.36	1.13	0.06	0.19	1.70

MP= Messzeitpunkt

Tabelle 124 Korrelationskoeffizienten

	Kennwerte	**1 MP**		**2 MP**	
		Musterlösung	Gruppenmodell	Musterlösung	Gruppenmodell
Struktur	Surface Matching	0.21	0.22	0.04	0.23
	Graphical Matching	0.06	0.14	0.11	0.16
	Structural Matching	0.19	0.19	0.14	0.14
	Gamma Matching	0.24	0.24	0.23	0.23
Semantik	Concept Matching	0.33 (*)	0.17	0.22	0.15
	Propositional Matching	0.26	0.05	0.28	- 0.10
	Balanced Semantic Matching	0.26	0.01	0.28	- 0.13

MP= Messzeitpunkt

D Musterlösungen

Musterlösungen der ersten Hochschulstudie

Abbildung 3 T-MITOCAR Modell der Musterlösung zum selbstregulierten Lernen

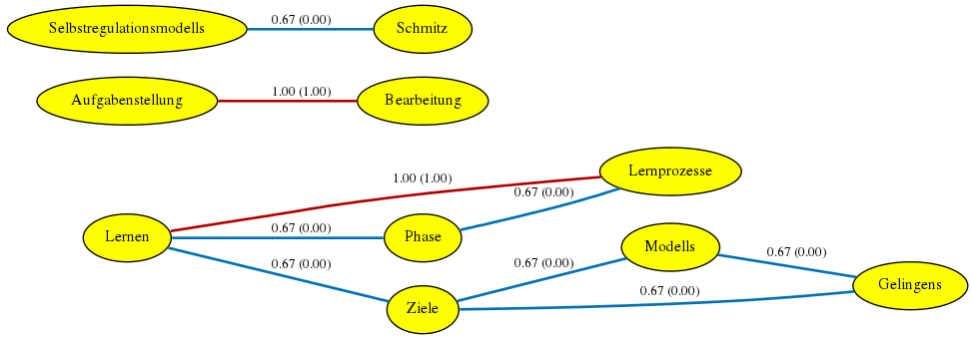


Abbildung 4 T-MITOCAR Modell des Gesamtmodells zum selbstregulierten Lernen (ohne Stichpunkte/GMO)

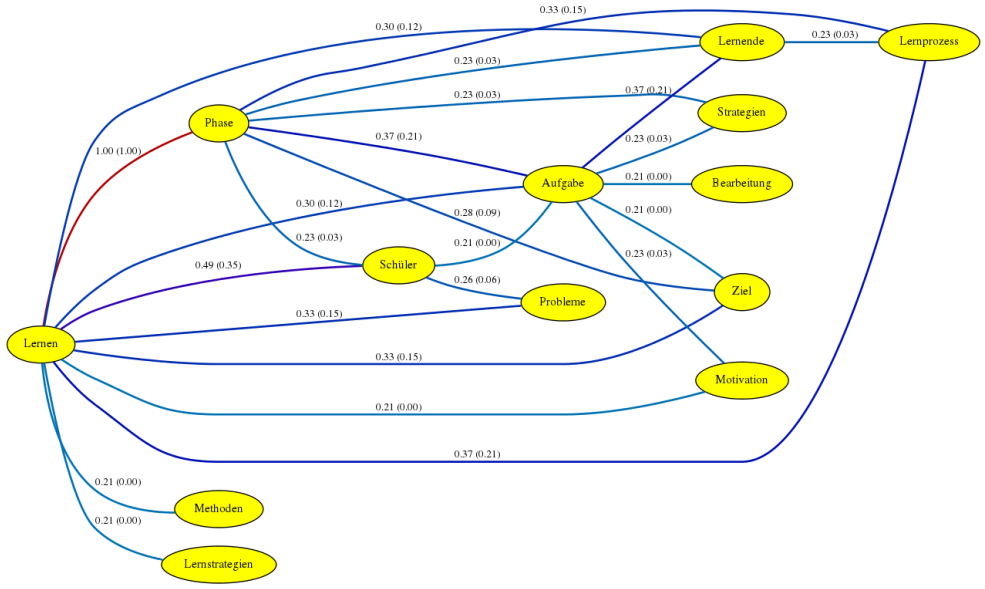


Abbildung 5 T-MITOCAR Modell des Lehrtextes zum selbstregulierten Lernen

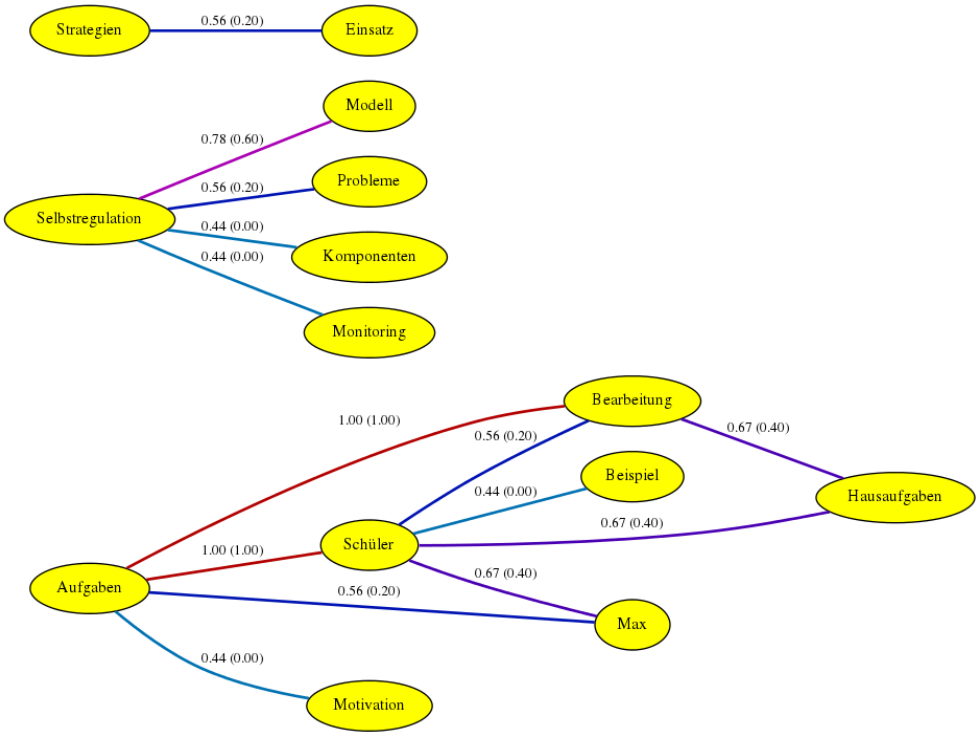


Abbildung 6 T-MITOCAR Modell der Musterlösung zur Metakognition

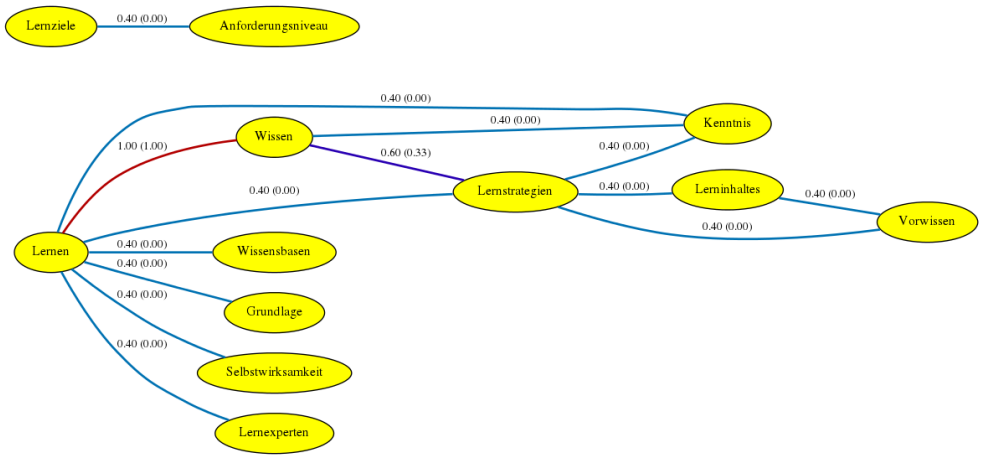


Abbildung 7 T-MITOCAR Modell des Gesamtmodells zur Metakognition (ohne Stichpunkte/GMO)

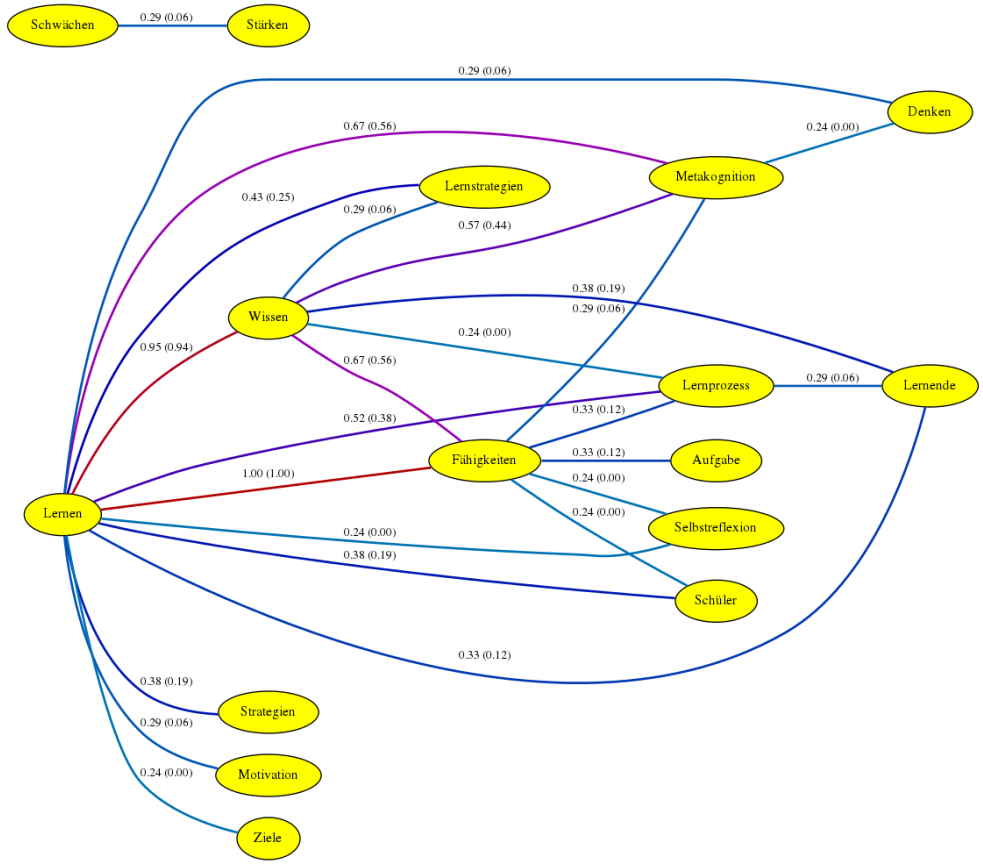
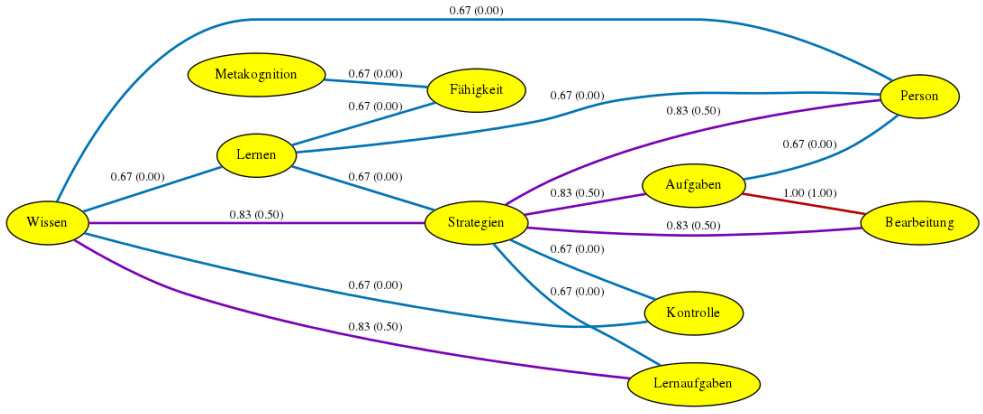


Abbildung 8 T-MITOCAR Modell des Lehrtextes zur Metakognition



Musterlösungen der zweiten Hochschulstudie

Abbildung 9 T-MITOCAR Modell des Gesamtmodells hinsichtlich des selbstregulierten Lernens

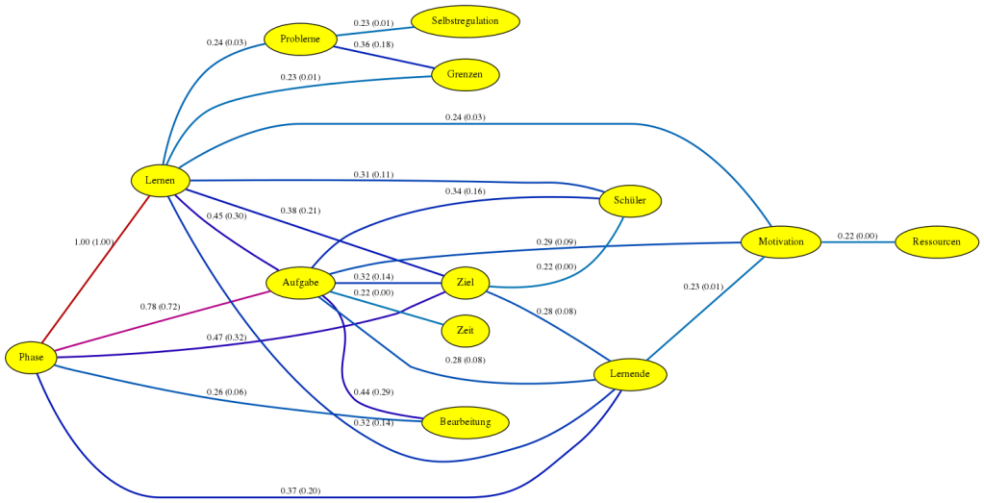


Abbildung 10 T-MITOCAR Modell der Musterlösung hinsichtlich der Lernstrategien

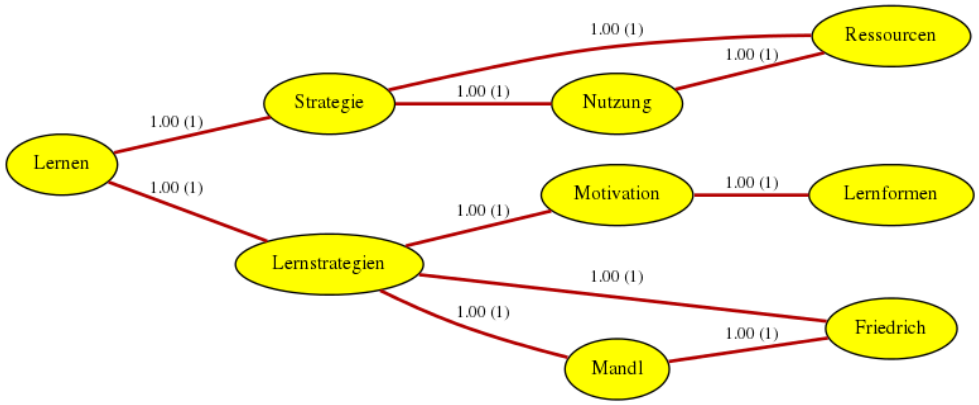


Abbildung 11 T-MITOCAR Modell des Gesamtmodells hinsichtlich der Lernstrategien

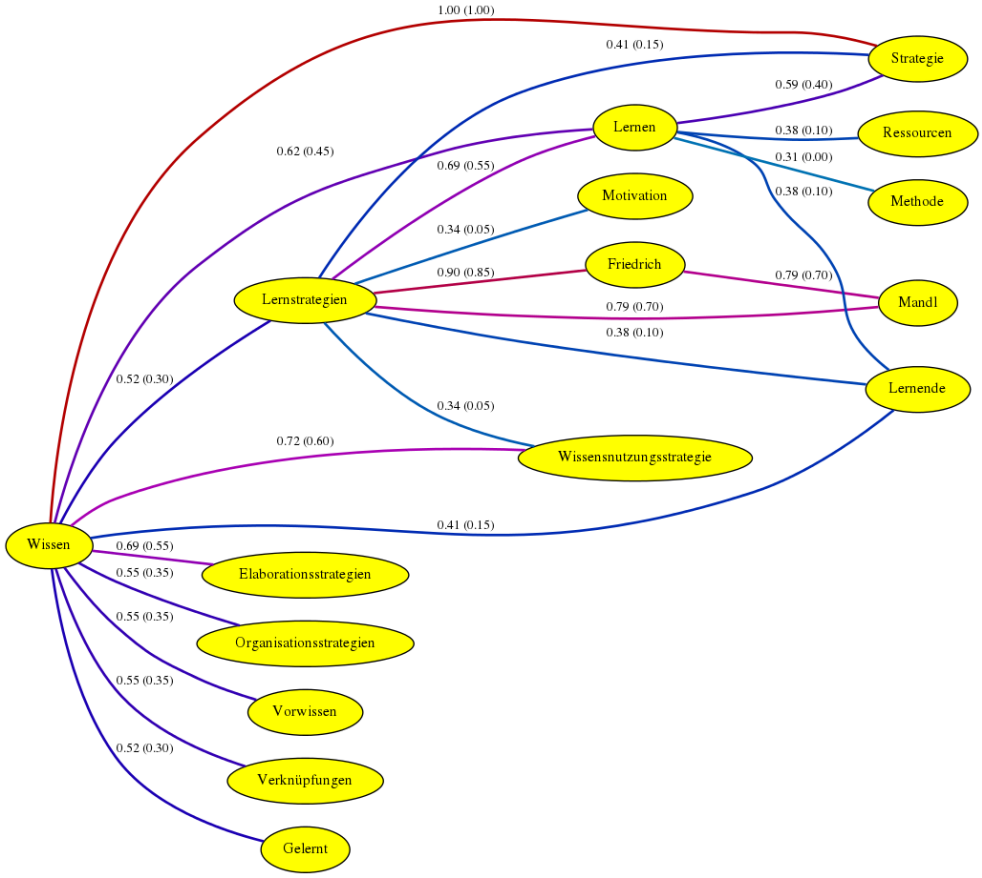
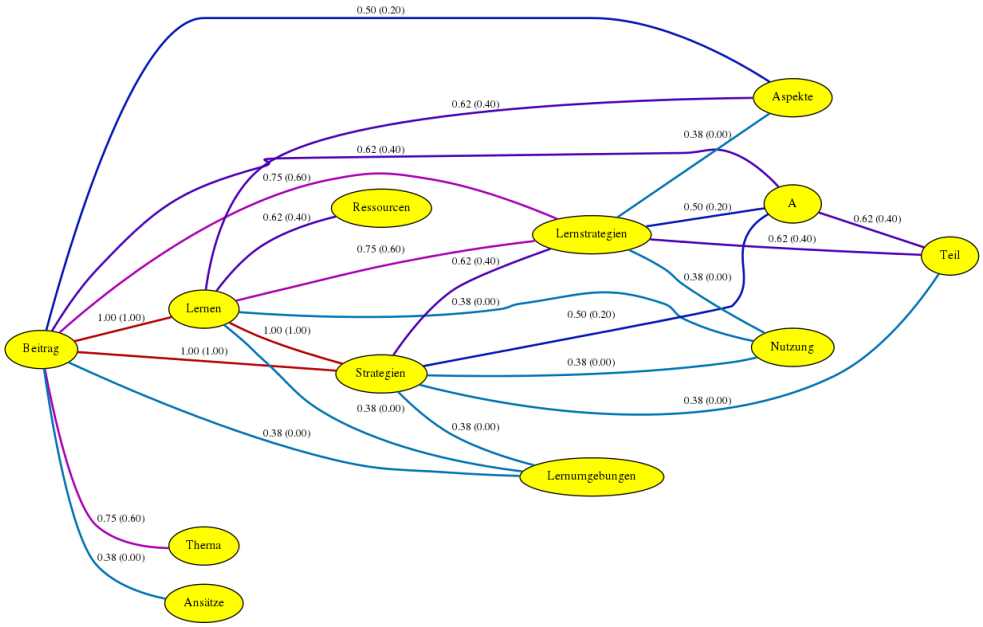


Abbildung 12 T-MITOCAR Modell des Lehrtextes hinsichtlich der Lernstrategien



Musterlösungen der Schulstudien

Abbildung 13 T-MITOCAR Modell der Musterlösung der ersten Lehrkraft im Unterrichtsfach Biologie

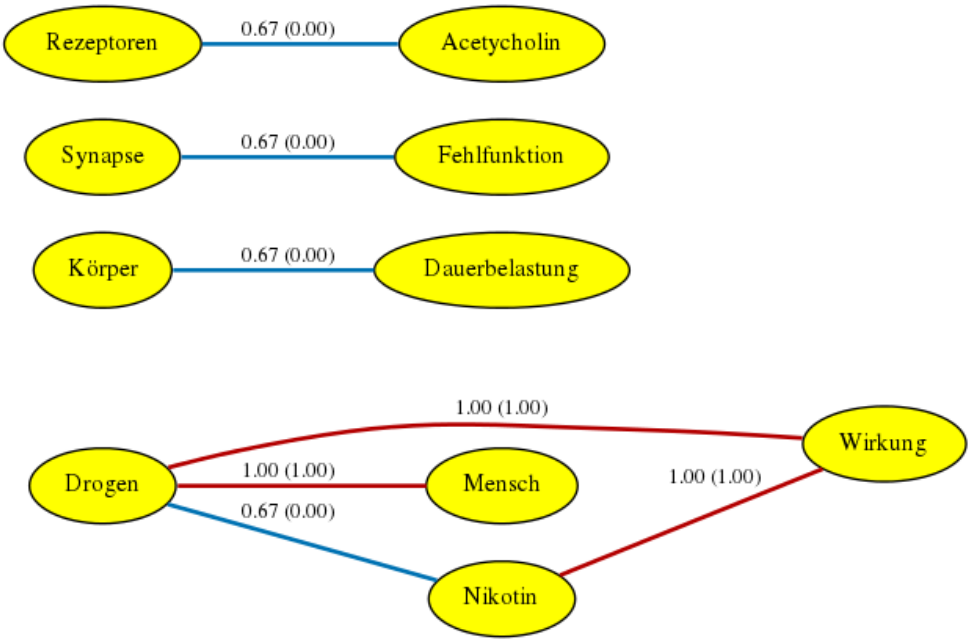


Abbildung 14 T-MITOCAR Modell des Gesamtmodells Biologie

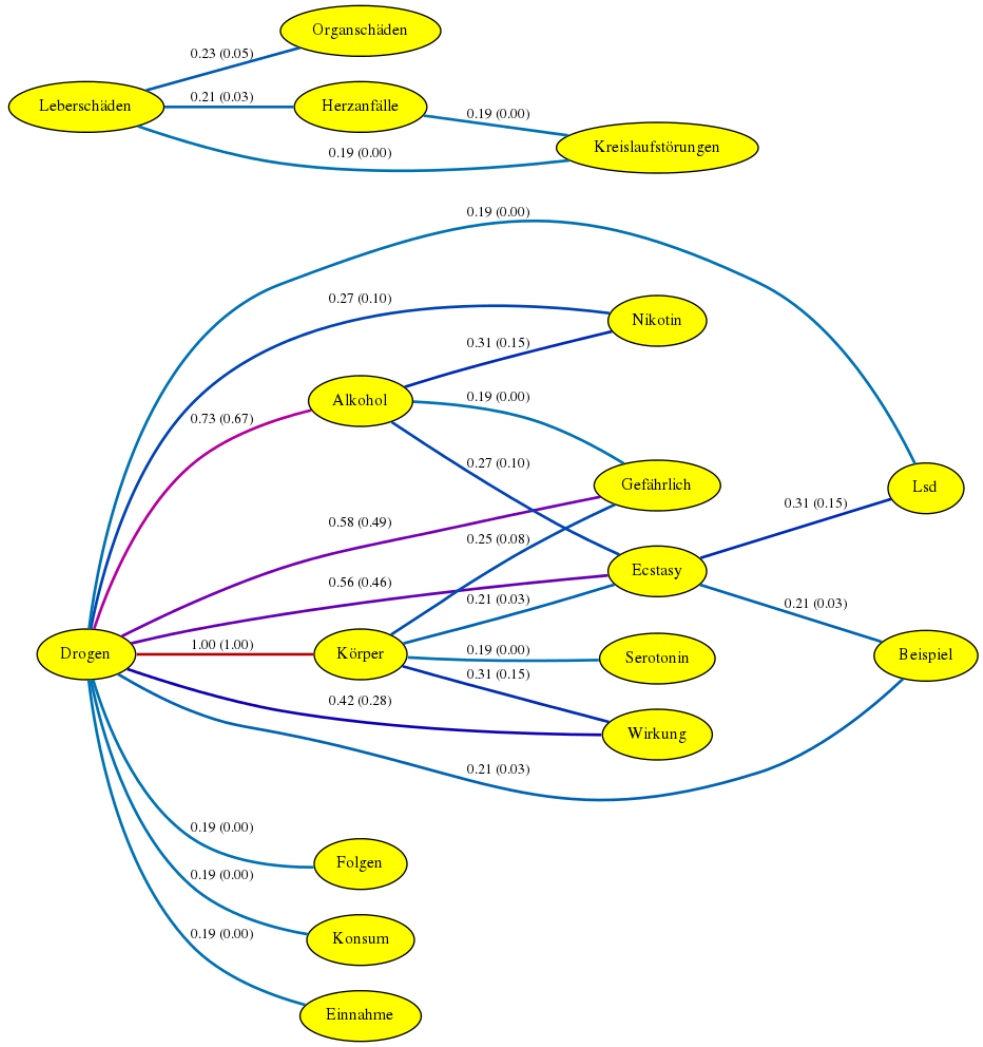


Abbildung 15 T-MITOCAR Modell der Musterlösung im Unterrichtsfach Deutsch

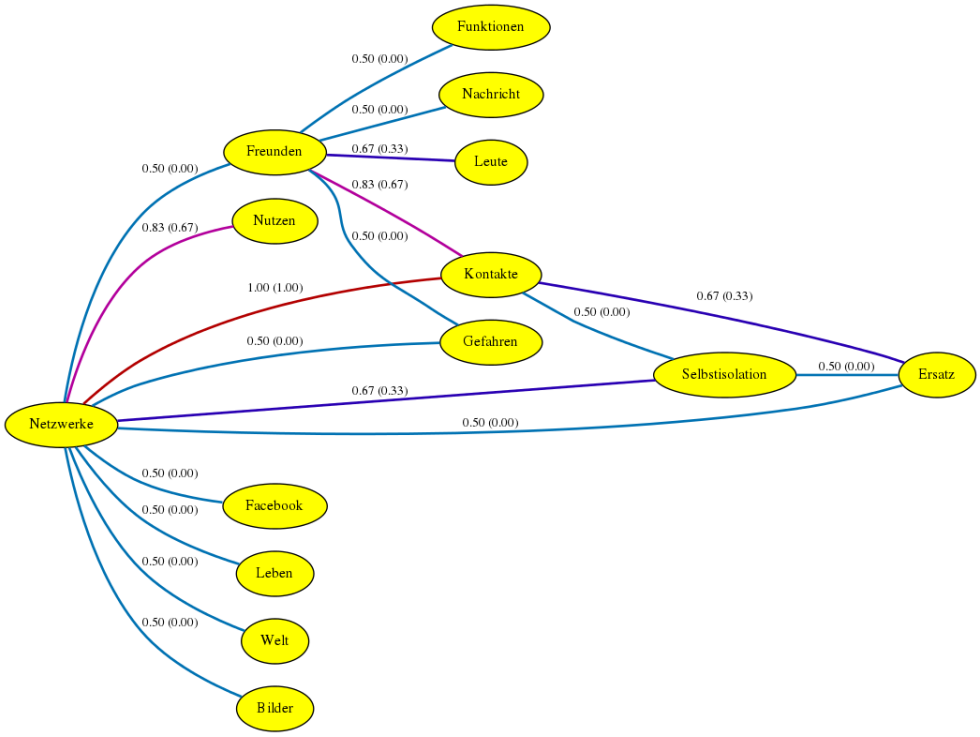


Abbildung 16 T-MITOCAR Modell des Gesamtmodells im Unterrichtsfach Deutsch

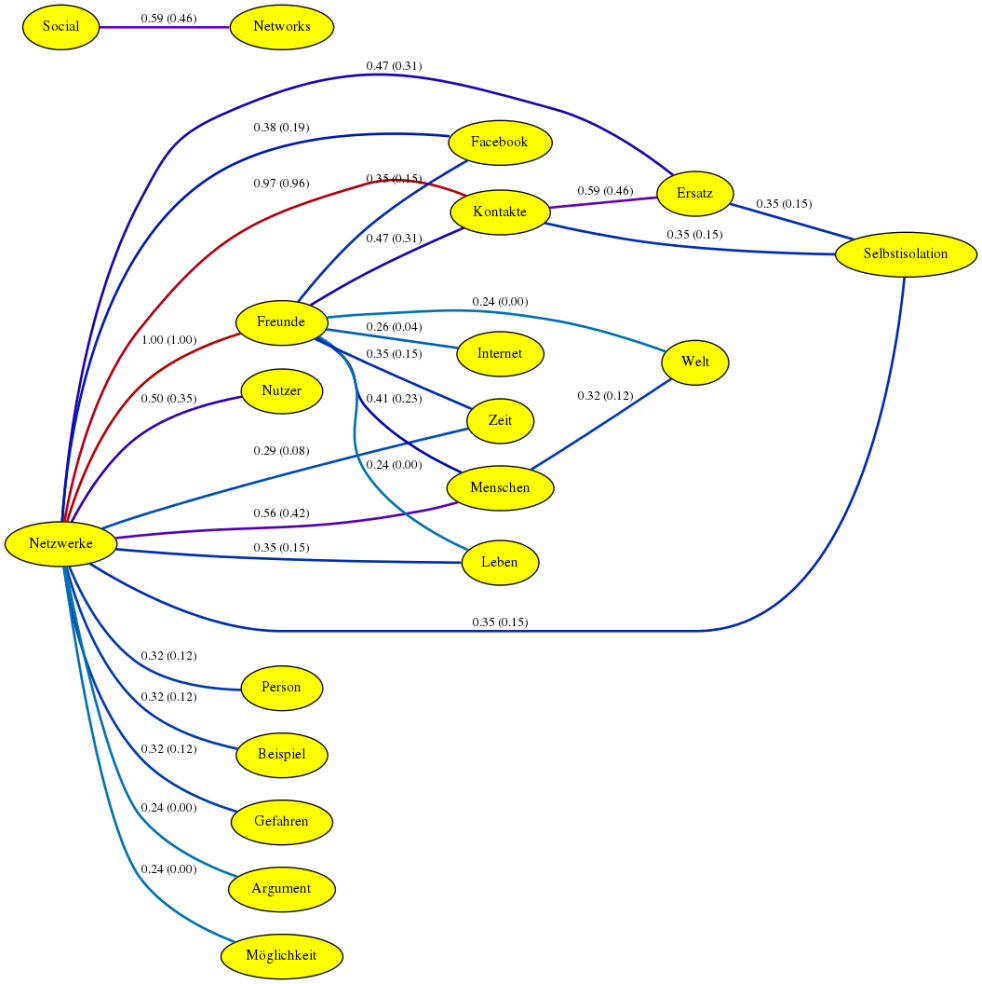


Abbildung 17 T-MITOCAR Modell der Musterlösung der ersten Lehrkraft im Unterrichtsfach Religion

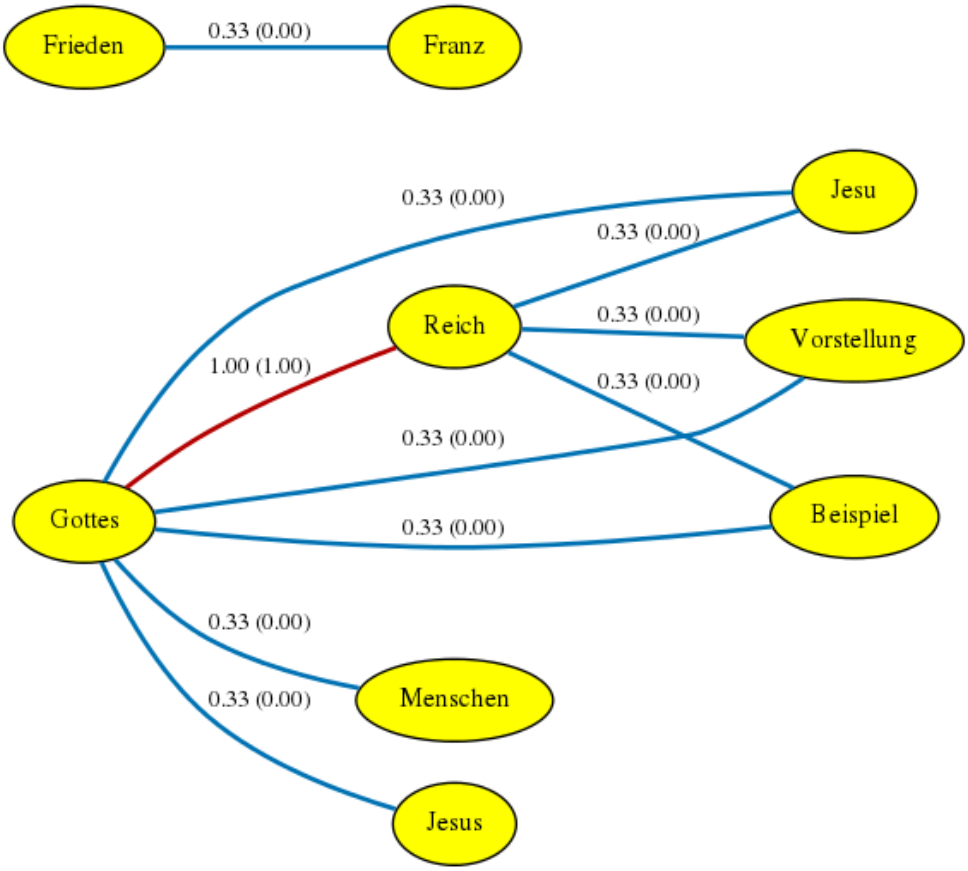


Abbildung 18 Gesamtmodell der ersten Teilstudie im Unterrichtsfach Religion

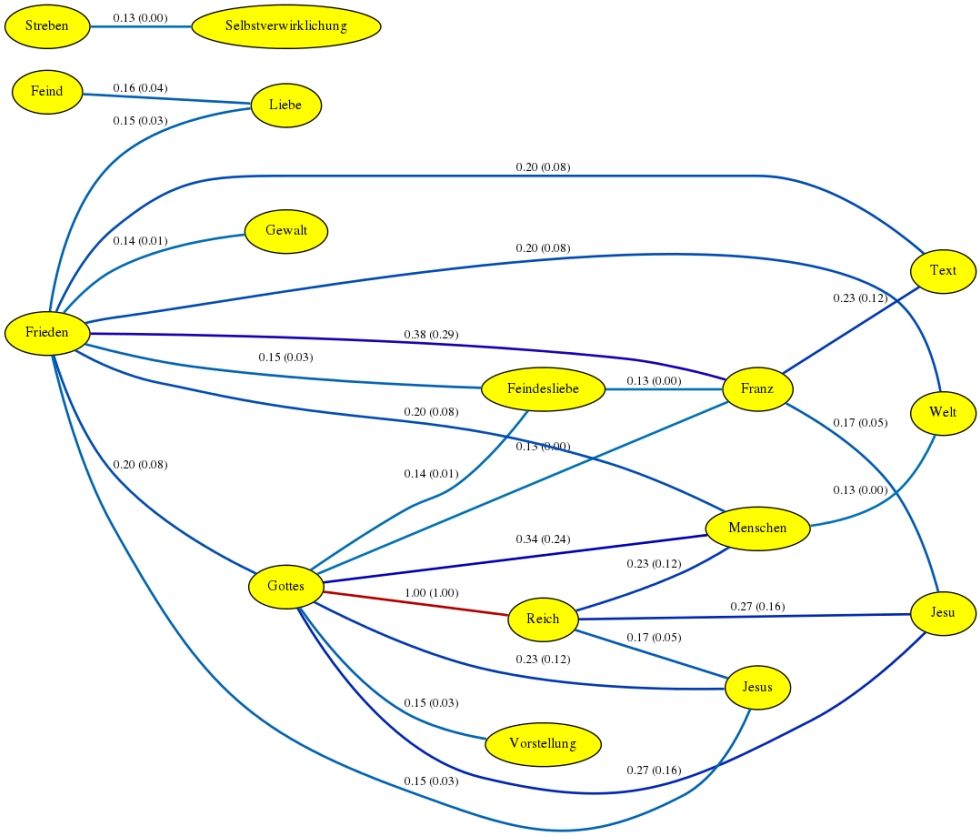


Abbildung 19 T-MITOCAR Modell der Musterlösung der zweiten Lehrkraft im Unterrichtsfach Religion

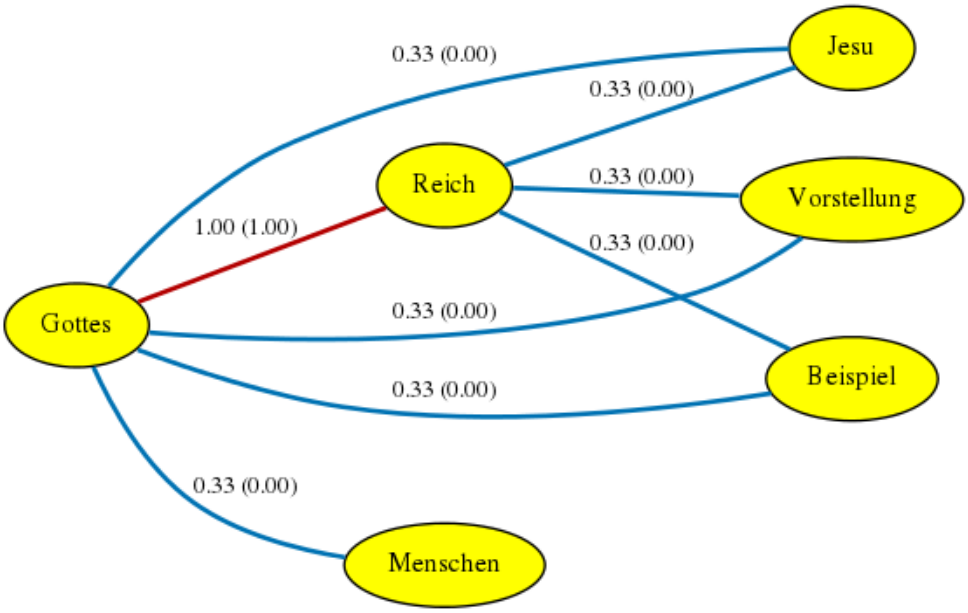


Abbildung 20 T-MITOCAR Modell des Gesamtmodells der zweiten Teilstudie im Unterrichtsfach Religion

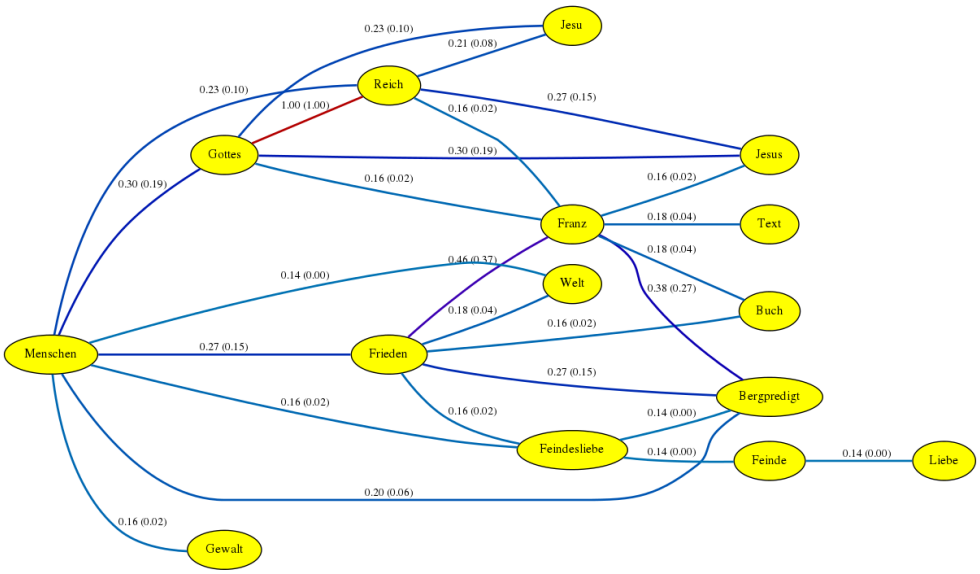


Abbildung 21 T-MITOCAR Modell der Musterlösung im Unterrichtsfach Kunst

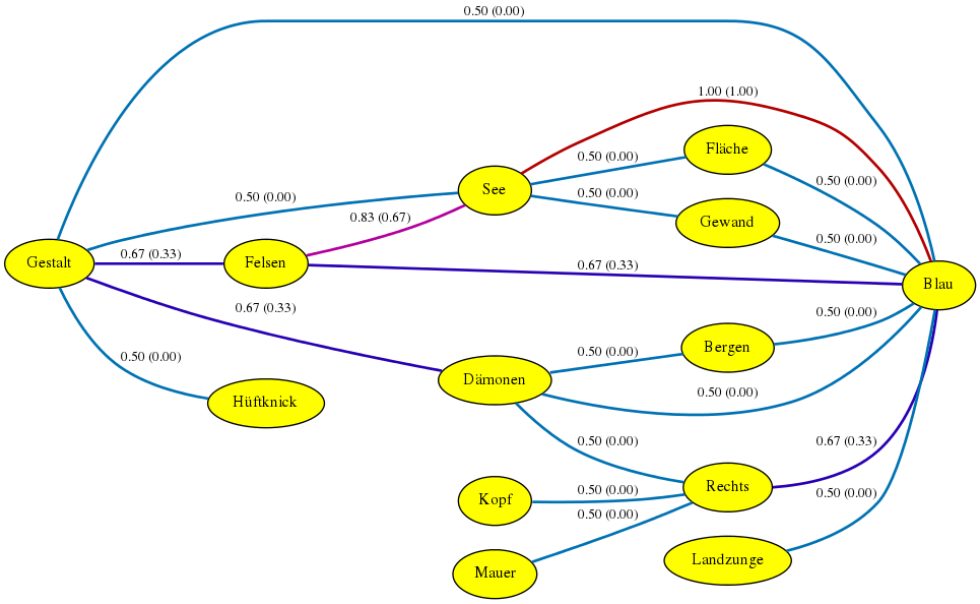
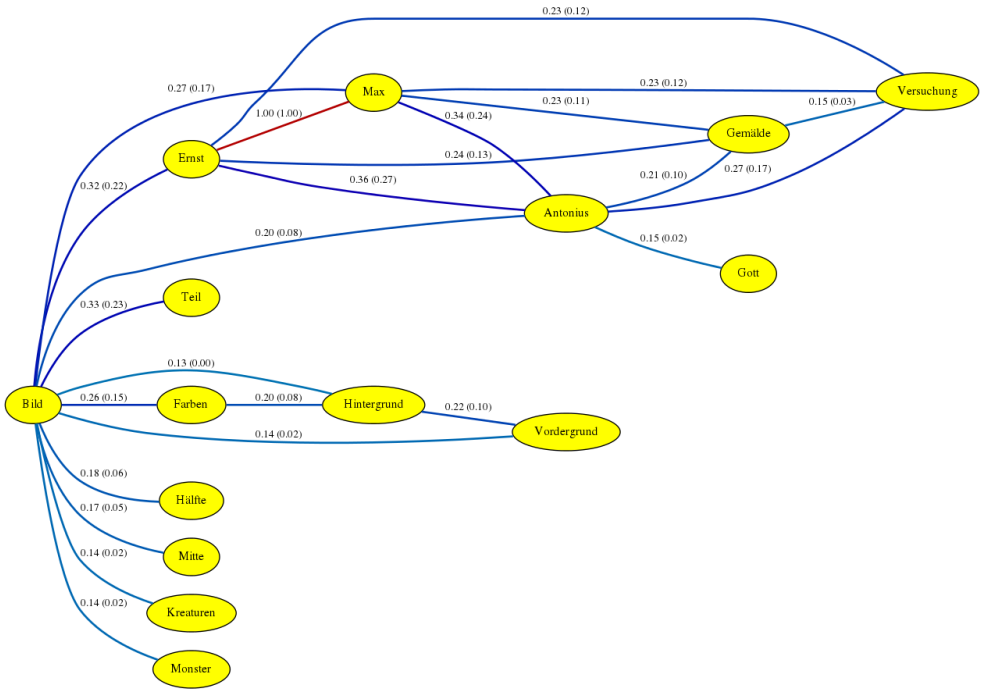


Abbildung 22 T-MITOCAR Modell des Gesamtmodells im Unterrichtsfach Kunst



E Interviews

Tabelle 125 Bewertungskriterien im Unterrichtsfach Deutsch (2. Messzeitpunkt)

13	L	Ok, ich bewerte textbasierte Schülerleistung nach Inhalt, Sprache und Form, grob
14		gesagt. Es gibt je nach Textform auch Unterschiede, aber das sind so die drei
15		Hauptkriterien.
16	I	Dann würde ich gerne gleich mit dir in deine Kriterien reingehen, die habe ich
17		gerade vor mir liegen. Da sind die drei Kriterien. Du hast sie fünfstufig unterteilt in:
18		sehr stark ausgeprägt oder sehr schwach ausgeprägt und mittelmäßig ausgeprägt. Zum
19		Beispiel: Beim ersten Kriterium „ Inhalt“. Wird in der Einleitung das Thema deutlich
20		benannt? Wenn ich die Texte mit deinen Kriterien noch mal bewerten würde, ab
21		wann könnte ich sagen, das ist total stark , oder es ist total schwach
22		Ausgeprägt? Was müsste ich suchen im Text oder finden, damit ich sagen kann: Ja?
23	L	Das ist ein Punkt, der schwierig ist für mich bei der Bewertung. Wir hatten
24		gesagt, wir verteilen Punkte für „Minus Minus“, „Minus Null“, „ Plus“, „Plus Plus“.
25		Ich habe das tatsächlich auch so gemacht. Normalerweise schaue ich mir eher an,
26		wo geht die Tendenz hin, wenn ich ganz viele Plus habe. Aber grundsätzlich
27		erst mal zu dem „Wo musst du da schauen?“ Wenn das Thema benannt wird, wenn
28		es für die Erörterung benannt wird, wäre das Null. Also dann wäre das
29		einfach vorhanden aber nicht besonders gut ausgestaltet. Wenn es so nur angestreift
30		wird, wäre das ein Minus, und wenn es gar nicht genannt wird ein Doppel-Minus. Und
31		Plus wäre, wenn es genannt wird und schon etwas besser formuliert ist.
32		Plus Plus wäre es, wenn es richtig gut formuliert ist, in diesem Fall noch
33		mit einer Definition versehen oder eine besonders gute Einleitung hat.
34	I	Das nächste Kriterium ist: Gibt es einen Überleitungssatz zum Hauptteil? Reicht das,
35		wenn da ein ganz kurzer Satz steht? Ist das sehr stark ausgeprägt, oder wie muss der
36		Überleitungssatz gefüllt sein?
37	L	Der Überleitungssatz, wenn der einfach vorhanden ist, aber nicht besonders ausführlich,
38		dann wäre das auch erst mal Null. Ist er gut formuliert, wäre es ein Plus
39		Wenn das sehr gut formuliert ist ein Plus Plus.
40	I	Und ob die Argumente stark oder schwach angeordnet sind? Wenn sie z. B. stark
41		angeordnet sind, geht es dann darum, dass jedes Argument, das der Schüler bringt,
42		sinnvoll ist? Ab wann ist es denn sinnvoll eingebracht?
43	L	Nein, wir haben das im Unterricht so besprochen, dass die Argumente, wenn wir eine
44		Pro-Kontra-Argumentation machen, dass sie erst mal die Kontra- Argumente
45		vorbringen. Und diese müssen von stark nach schwach geordnet sein. Dann kommt erst
46		das stärkste Kontra-Argument, dann werden diese immer schwächer und wie eine
47		Sanduhr. Danach fängt man auf der Pro-Seite mit dem schwächsten Argument an
48		und steigert sich bis zum stärksten. Also, dass der Aufbau so ist, dass der Leser am
49		Anfang denkt: Ja genau das spricht doch dagegen. Und dann wird er durch meine
50		Argumentation überzeugt, dass ich eigentlich Recht habe. Sodass am Schluss

das stärkste Argument durch meine Seite kommt.

I Die Argumente wurden durch Beispiele gestützt. Reicht es, wenn ein Argument mit einem Beispiel geschmückt ist oder jedes Argument?

L Nein, genau da geht es auch darum, weil sonst wäre das ja schwierig, zu bewerten. Hier Geht es um das Allgemeine. Wenn jemand immer gute Argumente benutzt hat, dann wäre es Plus Plus. Wenn jemand meistens gute Argumente benutzt hat wäre es nur Plus. Wenn Argumente da sind, jedoch Beispiele da sind, die nicht besonders überzeugend sind, dann wäre es Null. Minus wäre: Er hat es auch mal vergessen; Minus Minus: Er hat eigentlich kaum oder nie Beispiele benutzt.

I Ist schon eine starke Gewichtung von sinnvoll angeordnet, sinnvoll ausgewählt und sinnvoll durch Beispiele gestützt. Dann enthält der Schluss eine persönliche Meinung, eine weiterführende Fragestellung? Ab wann ist das Merkmal sehr stark ausgeprägt?

L Wenn jemand - das passiert oft bei Schülern - im Schlussteil einfach alles noch mal wiederholt, was er vorher schon genannt hat. Das hieße für mich, es gibt einen Schluss aber er ist nicht besonders gut gelungen. Sehr stark ausgeprägt wäre, wenn jemand wirklich noch weiter denkt. Er zum Beispiel noch seine persönliche Meinung einbringt und dann noch weiter denkt. Mir fällt jetzt gerade kein Beispiel ein. Halt einfach noch so eine offene Frage formuliert, in welche Richtung man noch weiter denkt, oder welches Problem generell noch offen bleibt, nachdem man schon alles andere betrachtet hat. Also das, was noch über den Text hinausgeht.

I Das erste sprachliche Kriterium ist: Wurden die Argumente mit unterschiedlichen Überleitungen eingeleitet. Wie würde ich hier vorgehen, wenn ich das erneut bewerten würde?

L Wir haben oft verschiedene Überleitungen im Unterricht besprochen. Z. B.: des Weiteren, außerdem, über dies, ein nächster Punkt ist. Das wäre im Prinzip Null, wenn jemand häufig „außerdem“ geschrieben hat. Das wären null Punkte, bzw. würde das schon ein Minus geben, wenn man immer das gleich benutzt. Plus Plus wäre, wenn jemand wirklich jedes Argument mit einer neuen Überleitung einleitet und das inhaltlich klar gegliedert ist, so dass dem Leser klar ist, wann immer das nächste Argument beginnt.

I Das nächste Kriterium: Ist der sprachliche Ausdruck angemessen?

L Da geht es darum, formuliert jemand sehr umgangssprachlich. Formuliert er sachlich und gut. Da wäre für mich Plus Plus, wenn jemand wirklich sehr gute Sätze baut, auch Haupt- und Nebensätze. Nicht nur parataktisch, sondern mit verschiedenen Konjunktionen und sich einfach inhaltlich gut ausdrückt. Ich weiß nicht, ob du die Aufsätze gelesen hast, oder wer das gelesen hat oder noch liest. Auf jeden Fall drücken sich manche Schüler sehr umgangssprachlich aus, und das wäre dann auch mal ein Minus Minus.

I Ok. Differenzierst du wenn, sich ein Schüler einmal umgangssprachlich ausdrückt und sonst sehr gut?

L Das wäre dann nicht Minus Minus. Ich würde sagen, das wäre dann Null. Also Null wäre für mich: Es ist einigermaßen, bzw. es ist in Ordnung, aber es ist nicht besonders gut formuliert. Und Minus wäre schon, wenn man immer mal wieder Sachen formuliert, die nicht besonders, stilistisch, nicht besonders gut sind. Plus, wenn jemand sehr gut formuliert. Plus Plus oder Plus, wenn es fast durchgängig gut formuliert ist, und Plus Plus wäre dann durchgängig gut formuliert.

I Das nächste Kriterium: sprachlich. Sind Rechtschreibung Grammatik und Zeichensetzung weitgehend in Ordnung? Was muss ich da finden oder suchen?

L Da habe ich eine bestimmte Zahl gesetzt. Ich glaube ab drei, bei Rechtschreibung Zeichensetzung und Grammatikfehlern, ist es Minus.

I Jeweils oder drei insgesamt?

L Jeweils. Also wer drei von jeder Form macht, der ist bei Minus. Wenn das dann noch doppelt so viele sind - mehr als sechs jeweils - dann ist es Minus Minus. Wenn es weniger als drei sind, ist es Null. Plus Plus wäre kein einziger Fehler oder vielleicht ein Fehler. Plus wäre meinetwegen zwei Rechtschreibungsfehler, ein Grammatikfehler.

I Nächstes Kriterium: Form. Da steht: „Lässt eine dreigliedrige Struktur eindeutig erkennen bei Einleitung Hauptteil Schluss.“ Ab wann gibt es da null Punkte?

L Da habe ich wirklich geguckt, ist es deutlich abgesetzt. Ist zwischen Einleitung Hauptteil und Schluss jeweils eine Zeile frei. Nur dann gab es Plus, beziehungsweise Plus Plus. Und Null gab es bei mir immer, wenn man schon irgendwie klar war, wann ein neuer Abschnitt anfängt aber nicht diese Zeile frei war und Minus entsprechend, wenn gar kein Übergang sichtbar war. Formal also, wenn irgendwann der Hauptteil anfängt aber man sieht optisch gar nicht, wo er beginnt. Und Minus Minus wäre, wenn der ganze Text komplett durchgeschrieben wäre.

I Ok., danke schön. Die nächste Frage: Erklär mal, wie du vorgegangen bist von der Aufgabenentwicklung bis zur Bewertung.

L Direkt die Aufgabe, die für die Klausur gestellt wurde?

I Ja.

L Wir haben im Unterricht dieses Thema bearbeitet, und die Schüler haben zu dem „sozialen Netzwerk“ viele Texte gelesen. Dann habe ich versucht aus diesen ganzen Arbeiten, die im Unterricht stattgefunden haben eine Fragestellung zu formulieren, die das mit einbezieht, beziehungsweise berücksichtigt, was da gearbeitet wurde. Das war die Aufgabenstellung. Die hat sich aus dem Unterricht ergeben. Die Bewertung haben wir in dem Fall nicht zusammen entwickelt. Das mache ich manchmal auch, aber in dem Fall habe ich ihnen beim Probeaufsatz diese Kriterien schon gegeben, damit sie wussten, was wichtig ist. Diese spiegeln im Prinzip wieder, wie wir das Thema erarbeitet haben. Wir haben eigentlich jeden, vor allem die Inhaltspunkte, aber auch die Sprachpunkte immer im Unterricht thematisiert. Also, dass das Thema genannt werden muss in der Einleitung, dass es eben diesen Überleitungssatz gibt, wie man die Argumente anordnen muss, dass man Beispiele

wählen muss, dass man die Argumente sinnvoll wählt, wie man den Schluss gestaltet. Auch sprachlich haben wir gesammelt: Wie kann man verschieden einleiten.

I Wie bist du dann zur Benotung gekommen mit den Kriterien, auch beim Wiederholten?

L Beim Wiederholten habe ich es wirklich nur mit Zahlen versehen.

I Beim Wiederholten meine ich die zweite Bewertung.

L Ja, genau.

I Genau.

L Da habe ich das mal mit Zahlen gemacht. Also 1,2,3,4,5 oben für die Kästchen. Form ist mir jetzt nur halb so wichtig wie die Inhalt und Sprache. Deswegen habe ich da dann immer nur 0,5 Punkte pro Kästchen gegeben, und das habe ich dann umgesetzt in ein Bewertungsschema. Das kann ich dir vielleicht auch noch schicken genau.

I Ok.

L Wo ich dann ablesen kann, für so und so viele Punkte gibt es eine Eins, dann eine Eins Minus und so weiter. Also absteigend und dann habe ich nachher die Noten verteilt. Diesmal fand ich es sehr schwierig. Normalerweise benutze ich dieses Kriterienraster als Rückmeldung für die Schüler, damit sie sehen, wo muss ich noch arbeiten, und das gibt mir im Prinzip nur so eine Richtung vor, was dann schlussendlich für eine Note dabei herauskommt. So musste ich mich wirklich festlegen, jedes Mal genau sagen, ja es diesmal jetzt genau Plus und diesmal genau Plus Plus, und das fand ich relativ schwierig.

I Wobei du das vorher auch eingetragen hast oder?

L Ja, doch nicht wirklich mit Punkten. Da habe ich gesagt: Ok, der liegt eher im Plus- und Plus Plus-Bereich, also muss es etwas zwischen Eins und Zwei sein oder jemand der im Nuller-Bereich ist, das ist komplett eine Drei. Ich habe es nicht so ganz genau gemacht.

I Welche Bearbeitungszeit gibst du Schülern, die in bestimmten Bereichen ein Defizit haben, beispielsweise Lese-Rechtschreibschwäche und auch Sprachschwierigkeiten?

L Ich habe in der Klasse niemand, aber hatte in einer anderen Klasse eine Schülerin, und sie hat immer eine halbe Stunde mehr bekommen pro Klassenarbeit. Die anderen haben anderthalb Stunden Zeit und sie eine halbe Stunde mehr.

I Das hast du mit ihr individuell vereinbart?

L Sie hat das mit der Beratungslehrerin vorher besprochen und die hat es mir gesagt.

I Welche Hinweise gibst du den Schülern beim Austeilen der Klausuren und wie reagierst du auf Nachfragen? Also ich meine beim Austeilen der Klausur, wenn die noch geschrieben werden?

L Ich lass die Schüler einmal den Text oder die Aufgabe durchlesen und sage: Es können gleich Fragen gestellt werden. Danach beantworte ich keine Fragen mehr, es sei Denn, es gibt jetzt irgendetwas, wo ich merke, da wissen wirklich alle nicht ganz Genau, was sie machen sollen. Dann sage ich es aber noch mal für alle von vorne.

I Und wenn inhaltliche Fragen gestellt werden, z. B., dass jemand fragt: Was kommt da noch mal hin?

L Nein, das sage ich nicht. Das müssen sie schon selber wissen.

I Schreiben deine Schüler beim Bearbeiten der Klausuren unter gleichen Rahmenbedingungen?

L Ich denke, die sind schon gleich für alle.

I Und woran machst du das fest?

L Na gut, ich kann nicht sehen, wie die jeweils persönlich gerade drauf sind an dem Tag. Aber das Klassenzimmer ist - schätz ich mal - überall gleich temperiert, und der Platz ebenso. Jedenfalls den, den die einzelnen Schüler haben. Davon würde Ich es jetzt abhängig machen. Und da würde ich schon sagen, es ist eigentlich für alle gleich.

I Welche Bewertungskriterien, kennen deine Schüler, bevor die Klausuren geschrieben werden?

L Genau dieses Schema.

I Erzähl mal, wie deine Bewertungsmaßstäbe zusammengekommen sind.

L Maßstäbe heißt jetzt was?

I Deine Bewertungskriterien.

L In dem Fall habe ich sie vorgegeben aus dem, was wir erarbeitet haben. Das mache ich nicht immer so, doch idealerweise sollte man sie mit den Schülern erarbeiten.

I Und tauscht du dich auch mit deinen Kollegen über Kriterien aus?

L Ja.

I Hast du schon mal versucht mehrere Bewerter heranzuziehen, beispielsweise in schwierigen Fällen? Und wie bist du dabei vorgegangen?

L Ja, habe ich schon versucht. Dann habe ich die Klausur kopiert, dem Kollegen gegeben und gebeten, dass er sie korrigiert, quasi wie eine Zweitkorrektur - einfach mit seinen Zeichen am Rand - und dann mir sagt, was er denkt, was das für eine Note sein könnte.

I Und was ist dann später herausgekommen?

L Eigentlich das Gleiche wie das, was ich mir vorgestellt hatte.

I Welche Inhalte erfassen deine Klausuren?

L Die Frage verstehe ich nicht, was ist damit gemeint?

I Wie gehst du vor, wenn du die Klausur entwickelst? Darauf bist du vorhin schon ein bisschen eingegangen. Welche Inhalte aus dem Unterricht werden da erfasst?

L Jetzt bei dem konkreten Fall oder allgemein?

I Bei dem konkreten Fall.

L Im Prinzip wurde genau das abgefragt, was wir auch behandelt haben. Sowohl vom Thema her „soziale Netzwerke“ als auch Formal haben wir genau die Sachen vorher geübt.

I Wie minimierst du Bewertungsfehler?

L Wie?

I Wie du Bewertungsfehler minimierst?
L Ich habe das jetzt mal versucht ‘nen bisschen nach der Schulung. Also wirklich so Pausen zu machen. Und hatte ja jetzt auch quasi die Arbeiten ohne Namen, um nicht zu überlegen, welcher Schüler das jetzt war. Und das so eher am Stück zu machen, wobei das jetzt nicht immer so gut funktioniert hat. Also ich hab dann doch an zwei Tagen insgesamt korrigiert und dann merkt man irgendwie doch, am zweiten Tag ist man doch irgendwie ein bisschen anders drauf wie am ersten.
Also ja man kann zusammenfassend sagen: Ich hab versucht insgesamt, alle gemeinsam, zeitnah zu korrigieren. Also hintereinander und dann aber auch Pause zu machen, wenn ich gemerkt habe jetzt werde ich irgendwie müde und kann mich nicht mehr konzentrieren und kreuze nur noch irgendwas an.
I Welchen Eindruck hast von einer Schülerleistung, wenn du unmittelbar davor eine sehr gute oder auch eine sehr schlechte Klausur bewertet hast?
L Wenn ich gerade eine sehr gute hatte, dann wird die nächste eher kritisch angeschaut, schätze ich mal. Und wenn ich was Schlechtes gelesen habe, dann erscheint die nächste schon erst mal besser.
I Hast du den Eindruck, dass deine Kollegen ihre Schüler strenger oder milder bewerten als du und woran machst du das fest?
L Ich denke es ist unterschiedlich. Es gibt Kollegen, die sehr viel strenger bewerten als ich und welche, die ähnlich bewerten wie ich. Und ich mach’s fest an dem, indem ich im Verzeichnis, wo wir die Noten eintragen müssen leicht sehen kann, wie die in anderen Hauptfächern benoten oder auch in andern Klassen.
I Hat deiner Meinung nach das Schriftbild einen Einfluss auf die Bewertung und wenn ja welchen?
L Ich denke schon, wenn’s auch vielleicht nur unterbewusst ist, aber ich denke schon; wenn jemand ordentlich schreibt und zum Beispiel auch gut gliedert, dass ich da tatsächlich dazu tendiere besser zu bewerten.
I Ok.
L Aber man kann es nicht so allgemein sagen, weil es gibt auch gute Schüler die eine Sauklaue haben (auf Deutsch gesagt) und da sehe ich auch manchmal über die Schrift hinweg.
I Wenn ein Schüler dir sympathisch oder auch unsympathisch ist, welchen Einfluss hat das auf deine Bewertung?
L Also, ich versuche eigentlich da nicht subjektiv zu bewerten. Das kommt dann vielleicht bei so was wie ich grad eben gesagt habe – wenn jetzt ein Schüler der sehr gut mitmacht im Unterricht, aber ‘ne unglaublich schreckliche Schrift hat – dann schaue ich in dem Fall vielleicht über die Schrift hinweg und konzentriere mich da eher auf den Inhalt.
I Und sonst nicht, wenn der im Unterricht nicht gut mitgemacht hat?
L Dann ist es vielleicht so unbewusst, dass man denkt ja das bestätigt sich, dass der da nicht mit macht und dann schreibt er auch nicht ordentlich. Das sind auch alles Sachen,

257		die versucht man natürlich nicht zu machen, aber das kann schon passieren gerade,
258		wenn man sowieso viel zu tun hat.
259	I	Welche Aufgabenniveaus berücksichtigst du in deiner Klausur, also welche
260		Schwierigkeitsgrade?
261	L	Es ist jetzt schwierig bei der Klausur, weil das nur eine Fragestellung war.
262		Normalerweise würde ich versuchen mindestens eine Frage, die sehr knifflig ist, noch
263		zu stellen und eine die sehr leicht ist, aber in dem Fall war jetzt tatsächlich nur eine
264		Fragestellung. Die war für alle gleich und konnten, meiner Meinung nach alle
265		lösen, die vorher mit gemacht und aufgepasst haben im Unterricht.
266	I	An welchem Maßstab orientierst du dich bei der Bewertung von Schülerleistung?
267	L	Damit sind jetzt nicht die Kriterien gemeint oder?
268	I	Der Maßstab, das können auch die Kriterien sein. Wenn du jetzt beispielsweise einen
269		Schüler hast, dem irgendwie noch ein Punkt fehlt aber der echt gut mitgemacht hat
270		im Unterricht. Bekommt der dann eine bessere Note?
271	L	Wenn das so ist, guck ich, ob ich noch irgendwo einen Punkt rausschinden kann, wenn
272		es wirklich an dem einen oder halben Punkt hängt.
273	I	Und wenn du merkst die ganze Klasse hat nicht eine Eins erreicht, gehst du dann mit
274		deinen Kriterien ein bisschen runter und sagst: Ok, ich bewerte weicher, so dass man
275		nicht meine ganzen Kriterien erfüllen muss, um eine Eins zu kriegen?
276	L	Also, es gibt meistens schon eine Eins, kommt aber auch vor, dass es keine Eins gibt
277		Also, es kommt auf die Klausur drauf an und auf die Klasse. Aber, wenn jetzt, die Eins
278		nicht so in unerreichbarer Ferne ist und man sie hätte erreichen können, dann gehe ich
279		eigentlich nicht runter.
280	I	Und wenn sie in unerreichbarer Ferne liegt?
281	L	Ja dann überdenke ich das noch mal, wenn ich jetzt sehe, wirklich keiner hat das
282		verstanden was ich wollte. Dann ändere ich’s auch noch mal ab.

Tabelle 126 Bewertungskriterien der zweiten Lehrkraft im Unterrichtsfach Religion (2. Messzeitpunkt)

1	I	Die erste Frage: nach welchen Kriterien bewertest du textbasierte Schülerleistungen?
2	L	Sind das jetzt dann wieder die gleichen Fragen?
3	I	Ja.
4	L	Ok. Nach welchen Kriterien bewerte ich textbasiert Schülerleistung? Also ich habe
5		einen Erwartungshorizont, an dem ich die Schülerleistung, die textbasierten ausrichte,
6		so in dieses Kriterium muss ich erst mal wieder reinkommen. Kriterien an sich für die
7		Bewertung. In diesem Fall habe ich eine Musterlösung benutzt, anhand derer ich die
9		Kriterien beurteile. Dann schau ich mir an, was wir im Unterricht gemacht haben und
10		prüfe anhand dessen ab - also von dem, was sie in der Lage sein sollten, das zu leisten.
11		Ja, vielleicht soweit erst mal.
12	I	Ich habe deine Kriterien bekommen, wenn ich jetzt deine Kriterien noch mal anlegen
13		würde und würde auch mal die Klausuren von deinen Schülern heranziehen und deine
14		Kriterien - ab wann wäre denn beispielsweise, die generelle Bewertung. Da hast du
15		erst mal Ausdruck und sprachliche Leistung. Auf was müsste ich da genau achten?
16		Also was muss genau im Text sein, damit ich sagen kann: Satzbau, Art der
17		Formulierung, Einbezug der Fachbegriffe sind sehr gelungen, gut gelungen
18		angemessen, teilweise gelungen, nicht gelungen? Ab wann kann ich denn sagen, dass
19		es nicht gelungen ist, dieses Kriterium Ausdruck, sprachliche Leistung?
20	L	Es wäre in dem Fall, wo sehr viel Umgangssprache benutzt wird, sehr „lapp“
21		formuliert. Ansonsten wäre es dann der Fall, wenn der überwiegende Teil der
22		Fachbegriffe, die notwendig sind - um bestimmte Sachen, die im Text auch
23		beschrieben sind- noch mal deutlich zu machen- nicht da sind. Und was noch dazu
24		kommt, wäre eine große Zahl an Rechtschreibfehlern oder Zeichensetzungsfehlern;
25		die hab ich aufgeschlüsselt per Quotient.
26	I	Aha, das heißt alle Unterpunkte müssten negativ, sag ich mal, belastet sein. Das heißt,
27		wenn einer viel umgangssprachlich schreibt, hat aber einen tollen guten Satzbau und
28		benutzt teilweise auch Fachbegriffe, dann wäre aber trotzdem das Kriterium nicht
29		erfüllt, weil er sich umgangssprachlich
30	L	Ja.
31	I	Wenn einer sich nicht umgangssprachlich ausdrückt aber die Fachbegriffe einbezieht
32		und das auch toll formulieren kann, der Satzbau aber grammatikalische Fehler aufweist,
33		wäre dann das ganze Kriterium nicht erfüllt?
34	L	Genau, ich habe versucht, insgesamt, natürlich nicht ganz akkurat aber schon das
35		gleich zu gewichten. Die einzelnen Teile Fachbegriffe zum Beispiel und die Art der
36		Formulierung. Das heißt, wenn jemand dann nicht gut formuliert hat, aber Fachbegriffe
37		genutzt hat, hat er schon dadurch einen großen Anteil richtig und ist deswegen dann
38		auch nicht im ungenügenden Bereich.
39	I	Und ist es bei Grammatik, Rechtschreib- und Zeichensetzung dann wieder ähnlich,
40		das heißt, wenn etwas grammatisch falsch ist, dann gibt’s auch hier für das ganze

Kriterium keinen Punkt? Oder wenn sie die Rechtschreib- und Zeichensetzung falsch ist?

L Wie es dann ist, das würd ich dann errechnen, Anzahl der Fehler auf die Anzahl der Wörter und je nachdem ein oder zwei Punkte abziehen.

I Ok, das geht dann Anzahl der Fehlerwörter, nicht Anzahl der Wörter im Text? Oder hast du die ganzen Wörter durchgezählt, weil, wenn einer viel Text schreibt, darf er ja vielleicht auch mehr Fehler machen?

L Ja, genau, da habe ich dann durchgezählt.

I Ok. Dann bei den inhaltlichen Kriterien, da hast du ja zum Beispiel Nennung von Autor, Quelle und Schwerpunkt des Textes „Franz Alt: Auszug aus ‚Frieden ist möglich'", 1986 um Beispiel Auslegung, Umsetzung der von Jesus geforderten Feindesliebe. Ab wann würde ich jetzt, wenn ich das noch mal bewerte, hier einen halben Punkt drauf geben?

L Also ich hab es im ersten Durchgang nicht so gemacht, dass ich tatsächlich einzelne Punkte darauf gegeben habe, sondern, so, dass ich mir den Text durchgelesen habe und dann erst mal grob eingeschätzt habe, in welchem Bereich das wohl ist. Und dann hab ich noch mal verglichen mit dem, was ich im Erwartungshorizont habe. Im zweiten Durchgang habe ich tatsächlich dann auch Punkte verteilt.

I Ob das ein halber oder ganzer Punkt ist, wenn ich den Text noch mal bewerten würde, worauf müsste ich achten, um diesen Punkt oder diesen halben Punkt vergeben zu können? Wenn jetzt der Schüler schreibt: Feindesliebe, irgendwas, Feindesliebe kommt vor, ist das dann ok?

L Also, wenn er es zu knapp geschrieben hat, dann würde ich die Hälfte der Punkte geben und wenn er es aber ausformuliert hat, dann die volle Punktzahl.

I Wenn einer nur „Franz Alt 1986" schreibt und nicht „Auszug aus ‚Frieden ist möglich'"?

L Genau, da in dem Fall, dadurch, dass das ja recht einfach ist, Autor und Quelle zu nennen, habe ich mich entschieden, dann den Schwerpunkt des Themas – also was der Schüler da hinschreibt, worum' s eigentlich geht - darauf ein Punkt zu geben und für Autor und Quelle dann jeweils einen halben.

I Zur Zusammenfassung der Hauptthesen. Zum Beispiel: Feindesliebe ist Alternative zur Kriegspolitik. Wonach müsste ich hier suchen? Reicht es, wenn der Schüler schreibt: Feindesliebe ist Krieg? Wie detailliert oder was muss genau im Text gegeben sein, um dann einen Punkt drauf geben zu können?

L Also hier müsste auf jeden Fall, inhaltlich mit drin stehen, dass Feindesliebe wirklich alternativ ist, dass es eine andere Möglichkeit ist, mit Konflikten umzugehen. An Begriffen meinst du jetzt speziell Begriffe, die drin stehen müssen?

I Nach was müsst ich suchen? Ok, dann weiß ich jetzt Feindesliebe, da muss man sagen, das ist eine Alternative Konfliktmanagement oder?

L Genau oder in dem speziellen Fall dann sogar Kriegspolitik.

I Und dann steht da zum Beispiel noch: Dadurch ist Heilung der Welt möglich. Wo

müsste ich danach dann im Text suchen?

L Da müsste dann deutlich werden, dass der Schüler versteht, dass eine positive Entwicklung der Welt möglich ist zumindest ein positiver Grundgedanke, dass die Welt sich verändern kann.

I Dann hast du noch Konjunktiv. Das heißt, konsequent meint: Es darf kein einziger Fehler drin sein.

L Genau, natürlich auf die Masse gesehen ist, wenn fünfzehn Mal Konjunktiv benutzt wird und einmal nicht, dann würde ich schon davon ausgehen, das ist ein konsequenter Gebrauch und dann gibt es mal einen Flüchtigkeitsfehler sozusagen. Aber wenn es einmal benutzt wurde am Anfang und dann nicht mehr, dann reicht mir das Beispiel nicht.

I Es muss sich durchziehen. So bist du dann auch in der zweiten und dritten Aufgabe vorgegangen. Das heißt, wenn ich da noch mal die Kriterien anlegen würde, jüdische Erwartung, Befreiung in der römischen Besatzungsmacht - also wenn der Schüler irgendwas über Besatzungsmacht schreibt, aber nicht sagt, dass es die römische Besatzungsmacht war - könnt ich ihm dann trotzdem noch den Punkt geben?

L Ja, ich würde wahrscheinlich dann leichte Abzüge machen, je nachdem wie er es geschrieben hat. Ob das durch den Text vorweg oder hinterher noch klar wird, dass es verstanden wurde. Das Problem war, wenn sie dann einfach geschrieben haben: „die römische Herrschaft", das fand ich auch etwas unkonkret, da würde ich dann Abzug machen.

I Dann, ja ‘n halben Punkt gegebenenfalls, und dann müssten jetzt nicht alle drei Punkte genannt sein. Oben eins von beiden. Und bei Zeloten: „wollen das Reich Gottes mit Gewalt schneller herbeizwingen“. Das reicht dann, wenn der Schüler schreibt: die Zeloten gehen gewaltsam vor oder?

L Ja, genau.

I Ok, und auch das Fazit gab noch einen halben Punkt, wenn der Schüler so schreibt: Jesus Vorstellung vom Reich Gottes widerspricht der jüdischen Erwartungshoffnung, dann würde es da auch einen halben Punkt geben? Würde das als Fazit ausreichen, da muss nicht alles gegeben sein sondern immer mindestens eins von den Punkten und das ist dann auch egal, ob das kurz gefasst ist oder detailliert. Hauptsache es ist irgendwie genannt.

L Genau, wenn es einem Fazit entspricht. Also noch mal so eine kurze Zusammenfassung dessen, was geschrieben wurde. Wenn plötzlich im letzten Satz noch mal was ganz anderes steht als davor, dann würde ich es nicht als Fazit gelten lassen.

I Dann erkläre mal, wie du vorgegangen bist von der Aufgabenentwicklung bis zur Bewertung.

L Also, ich habe erst mal geschaut, was wir im Unterricht alles gemacht haben, was ich ja jetzt, wenn ich’s noch mal machen würde, noch anders machen würde.

I Wie denn?

L Also jetzt würde ich mich die Tage mal hinsetzen und die nächste Klausur erst mal

konzipieren und dann darauf ausgehend meinen Unterricht aufbauen. Andersrum, jetzt hab ich's eher gemacht, auch in der Form, haben wir so weit gesichert, dass man es drannehmen kann, also nicht unbedingt irgendwelche Gruppenpräsentationen sondern schon Sachen, die gut an der Tafel standen. Zum Beispiel die, wo sie zumindest in der Lage waren, das gut mit aufzunehmen. Ich hab mir dann angeschaut inwieweit ich die drei Anforderungsbereiche mit aufgreifen kann und dann eben, dass ich wusste, ok, da nehme ich einen Text als Grundlage, wo sie zeigen sollen, dass sie ihn verstehen und auch zusammenfassen können. Im zweiten dann eben auch das mit einbringen können, also eher erläuternd werden, zu dem, was wir dann im Unterricht dazu gemacht haben. Das in Bezug zu Setzen und die dritte Aufgabe hat ja nicht so gut geklappt da noch mal einen Transfer zu leisten, wo sie dann eine eigene Interpretation noch mal aufschreiben. Von der Entwicklung her, hab ich auch geschaut, dass es ein Text war, der sprachlich angemessen ist, von dem her, was die Schüler verstehen können. Also keinen hochtheologischen Text und auch von der Länge angemessener Text, der in anderthalb Stunden gut zu schaffen ist. Durchführung: Die Klausur war ja auf 90 Minuten angelegt. Erst mal aufgeteilt, dann noch mal kurz die Aufgaben erklärt und was man dazu sagen muss. Darauf geachtet, dass sie nicht voneinander abschreiben können. Das heißt: Sie saßen weit auseinander, was bei 23 Schülern ganz gut geht in dem Raum, indem wir waren. Bei Fragen habe ich dann noch mal überlegt, ob ich da für alle noch mal was erkläre oder ob das jetzt einfach eine spezielle Frage war, die von einem kam und zu der ich dann noch mal einzeln vielleicht kurz was dazu gesagt hab. Oder aber es war dann wirklich, für viele eine Hilfe und schon eine Lösung, die ich hätte sagen müssen. Dann hab ich gesagt: Dazu kann ich, nichts sagen.

I Wie bist du da zur Bewertung gekommen? Zur Notengebung?

L Meinst du jetzt den ersten oder zweiten Durchgang?

I Den Zweiten.

L Den Zweiten. Genau, in dem Fall habe ich mir konkret doch noch mal überlegt, zu diesen einzelnen Punkten, die ich im Erwartungshorizont hatte, noch mal Bewertungspunkte zu geben. Das hatte ich bei dem ersten mir auch erst überlegt, habe das dann aber anders gemacht - hab ich dir ja schon erklärt. Und in dem Fall hab ich jetzt, tatsächlich überlegt, inwieweit die einzelnen Punkte gewichtet sind und wie wichtig mir ist, dass das drin steht. Dementsprechend habe ich das dann auch versucht, von den Punkten her zu verteilen. Jeweils habe ich immer drei Punkte für Sprache gegeben und zwölf Punkte für Inhalt. Beim Korrigieren habe ich das wieder so gemacht, dass ich Aufgabe für Aufgabe korrigiert habe. Also erst mal bei allen Aufgabe Eins, dann bei allen Aufgabe Zwei; habe auch versucht, schön entspannt zu sein, also nicht, völlig abgearbeitet bei der letzten dran zu sitzen, sondern möglichst Pausen einzubauen, um da möglichst in der gleichen Stimmung zu sein

I Im Urlaub?

L Ja, genau. Wobei, wenn man im Urlaub Klausuren korrigiert, ist man auch nicht in

so guter Stimmung. Und dann dementsprechend die Punkte mir angeschaut, die da waren und die ich auch erwartet hatte. Auch geschaut, wo vielleicht noch Zusätze sind, die ich selber nicht erwartet hatte aber die positiv in die Note eingehen. Dann habe ich überlegt, wie viel Punkte könnte man da noch als Zusatz geben, sodass - wenn beispielsweise bei Inhalt ein Punk nicht erreicht wurde, aber dafür zwei andere kleinere Sachen noch erwähnt wurden, die ich nicht erwartet hatte - dann der Punkt dann wieder ausgeglichen war.

I Und wie ist die Note dann zustande gekommen?

L Die Note ist dadurch zustande gekommen, dass ich jede Aufgabe, mit fünfzehn Punkten bewertet habe. Also mit diesen drei sprachlich, zwölf inhaltlich, hatte ich dann ja jeweils zwischen null und fünfzehn Punkten und habe aber die die beiden Aufgaben unterschiedlich gewichtet. Die erste Aufgabe in dem Fall mit 43 Prozent, die andere mit 57 Prozent. Das war bei drei Aufgaben dann noch anders, da hatte ich 30, 40, 30 Prozent und das hab ich jetzt umgerechnet auf zwei Aufgaben, sodass ich dann diese Gewichtung hatte. Das heißt ich habe die einzelnen Punkte zusammengezählt, die derjenige erreicht hat und daraus hat sich dann die Note ergeben.

I Das heißt Maximalpunktzahl war fünfzehn?

L Genau.

I Ok. Welche Bearbeitungszeit gibst du Schülern, die in bestimmten Bereichen ein ein Defizit haben, beispielsweise Lese-Rechtschreib-Schwäche oder auch Sprachschwierigkeiten?

L Also in dem Fall war es Oberstufe, wo die Lese-Rechtschreib-Schwäche nicht mehr gewertet wird. Ich hatte auch keinen in meinem, Kurs, wo das der Fall war, von daher hat sich das Problem jetzt nicht ergeben. Ich hatte einen Schüler drin, der noch nicht so gut Deutsch schreiben konnte, weil er vorher in der ausländischen Klasse war aber irgendwann gewechselt hat. Auch da habe ich, weil das Oberstufe ist, nicht noch extra Zeit gegeben.

I Welche Hinweise gibst du deinen Schülern beim Austeilen der Klausuren und wie reagierst du auf Nachfragen? Also ich meine beim Austeilen, wenn die Klausur geschrieben wird?

L Beim Austeilen ist so, dass ich erwarte, dass alle gleichzeitig anfangen. Das heißt ich drehe es erst noch mal um bis es alle haben, sodass dann die Zeit für alle gleich ist. Gebe dann die Hinweise noch mal, die immer auch dabei stehen, auf die Rechtschreibung zu achten, weil das halt auch mit rein-zählt. Sich die Aufgaben gut durchzulesen, auch da auf die Gewichtung zu achten, sodass sie wissen, bei der Aufgabe, die mehr zählt, müssen sie auch dementsprechend, ein bisschen mehr Zeit einplanen. Und ich gebe auch noch mal einen Hinweis auf die Gesamtzeit, die sie zum Schreiben haben. Dann erkläre ich noch mal kurz die Aufgaben. Wirklich nur kurz, weil eigentlich versuch ich es schon so zu formulieren, dass sie verständlich sind.

I Und wie reagierst du auf Nachfragen?

L Bei Nachfragen, manchmal ist es ja so, dass gleich am Anfang Nachfragen kommen,

die lasse ich erst mal laut stellen und sage was dazu oder aber auch nichts. Und sage: Wenn Fragen sind, komm ich rum. Und bei Einzelnachfragen habe ich schon gesagt, lass ich auf jeden Fall diese Frage auch zu und schau dann, ob es wirklich eine Einzelfrage ist, oder ob sie mehrmals auftaucht, und ich das komplett noch mal mit dem Kurs klären muss. Oder auch, dass ich an der Art der Frage merke, oh, da habe ich jetzt selber nicht drüber nachgedacht, das stimmt, das muss ich noch mal laut sagen.

I Und, welche Art der Antworten gibst du den Schülern? Auf welcher Ebene? Wenn ich jetzt sage: Ich weiß gerade nicht, was damit gemeint ist?

L Je nachdem, wenn ich merke, der Schüler weiß es nicht, weil er nicht mehr weiß, was wir dazu im Unterricht gemacht haben, dann helfe ich nicht. Wenn ich merke, es ist jetzt einfach wirklich nur eine Verständnisschwierigkeit, wie das dann insgesamt gemeint ist, dann formulier ich noch mal um und erklär es noch mal.

I Welchen Eindruck hast du? Schreiben deine Schüler beim Bearbeiten der Klausuren unter gleichen Rahmenbedingungen?

L Grundsätzlich habe ich den Eindruck, dass sie so weit unter gleichen Rahmenbedingungen schreiben, als sie natürlich Fähigkeiten mitbringen. Wenn sie Schwierigkeiten haben, sich, schriftlich auszudrücken, dann merk ich es schon. Dann sieht die Klausur dementsprechend aus. Ich habe eine dreier Gruppe im Kurs, die mündlich sehr gut sind, aber schriftlich dann doch eher bei den Schlechteren. Insofern sind natürlich die Rahmenbedingungen nicht ganz gleich, aber es wird auch gefordert einen schriftlichen Teil mit drin zu haben und deswegen ist das auch notwendig. Ansonsten versuche ich die Rahmenbedingungen soweit wie möglich für alle gleich zu gestalten.

I Welche Bewertungskriterien kennen deine Schüler bevor die Klausur geschrieben wird?

L An sich gehen wir immer die Aufgabentypen durch. All Anforderungsbereiche vor der Klausur, und da erkläre ich auch noch mal, worauf es mir ankommt. Das war schon die zweite Klausur mit diesem Kurs und da hab ich auch noch mal gesagt, sie sollen sich auf jeden Fall noch mal die erste Klausur angucken, was ich jeweils dazu geschrieben habe, wo so Schwachpunkte waren oder so, dass sie da auch die Möglichkeit haben noch mal zu wissen, worauf kommt es mir an; welche Kriterien sind wichtig. Das können sie daraus dann auch ganz gut ablesen.

I Und wissen die Schüler, ab wann es Punktabzug gibt oder stehen auch die Punkte an der Aufgabenstellung dran?

L Also die Punkte selbst stehen nicht dran, sondern nur die Gewichtung der einzelnen Aufgaben und an sich der Punktabzug. Also sie wissen schon, wenn sie eben nicht vollständig darauf antworten oder falsche Sachen mit rein bringen, dass es dann natürlich Punktabzug gibt. Oder wenn zum Beispiel bei einer Inhaltsangabe plötzlich schon interpretiert wird oder eine eigene Meinung reingebracht wird. Solche Dinge wissen sie dann schon.

I Dann erzähl mal, wie deine Bewertungsmaßstäbe zusammengekommen sind.

L Also in dem Fall hab ich entsprechend den Erwartungshorizont erstellt. Das heißt, ich habe noch mal geschaut, was können sie wissen, was geht aus dem Text hervor, den sie als Grundlage kriegen und habe darauf einen Erwartungshorizont erstellt plus eine Musterlösung. Die habe ich mit berücksichtigt, allerdings hat mir der Erwartungshorizont selbst mehr geholfen und da habe ich dann entsprechend Sprache und Inhalt getrennt. Sprache: drei Punkte, Inhalt: zwölf Punkte pro Aufgabe und habe da jeweils dann noch mal zum einen überlegt, was für Kriterien waren mir wichtig? Auch jeweils zu den einzelnen Aufgabentypen. Zum Beispiel, wenn ich schreibe, sie sollen in eigenen Worten zusammenfassen, dass das eben als Kriterium auch reinkommt und habe die einzelnen Punkte aufgelistet, die ich inhaltlich hören wollte.

I Tauschst du dich mit deinen Kollegen aus über die Bewertungsmaßstäbe oder macht das jeder so für sich?

L Im zweiten Durchgang, jetzt nicht. Das habe ich jetzt komplett alleine gemacht.

I Hast du schon mal versucht mehrere Bewerter heranzuziehen? Beispielsweise in schwierigen Fällen. Wie bist du dabei vorgegangen?

L In schwierigen Fällen habe ich auf jeden Fall mit Kollegen Rücksprache gehalten. Jetzt in diesem Fall, wie gesagt, nicht aber ansonsten schon. Sodass ich, wenn ich mir unsicher war, einfach noch mal einen Kollegen gefragt habe. Ich habe es dann so gemacht, dass er nicht wusste, welche Note ich gegeben habe. Sondern habe nur gefragt, ob er mal drüber guckt und mal überlegt, was er geben würde, und habe das in der Notenfindung berücksichtigt. Wenn ich zum Beispiel eine schlechte Punktzahl gegeben hätte und der Kollege eine gute, dann hätten wir uns darüber ausgetauscht - warum er denn denkt, dass das eine gute Leistung ist und ich gesagt, gesagt, warum es eine schlechte Leistung ist - und hätten dann gemeinsam geguckt, wo kann man sich denn da treffen.

I Und welche Inhalte erfassen deine Klausuren?

L Also die Inhalte sind bezogen auf das, was im Unterricht vorher gemacht wurde, in dem Fall immer semesterweise. Allerdings sage ich vor der Klausur noch mal welche Themen aus dem Unterrichtsinhalt drankommen, beziehungsweise welche sie sich auf jeden Fall angucken müssen und welche wegfallen. Wenn es dann zu viel war, was wir auch im Unterricht haben, dann schränke ich ein bisschen ein. Ansonsten kommen die Inhalte dran, die vorgegeben sind durch den Text. Also es ist ja letztendlich ein neuer Text, den sie kriegen, den sie erfassen müssen. Insoweit Unterrichtsinhalt und der neue Inhalt des Textes, den sie dann bekommen.

I Wie minimierst du Bewertungsfehler?

L Ich versuche immer möglichst außen vor zu lassen, wer das jetzt geschrieben hat und nicht darauf zu achten sondern am Erwartungshorizont abzugehen, was drin steht und was nicht, - möglichst objektiv. Um eine bessere Vergleichbarkeit zu haben, gehe ich Aufgabe für Aufgabe Durch. Jeweils bei jedem erst mal Aufgabe eins und, wie gesagt, ich versuche mehr Pausen einzubauen und auch die Schrift außen vor zu lassen.

Da ist die Gefahr, wenn es dann zu krakelig geschrieben ist, dass man schon denkt, das das kann auch nicht so gut sein. Auch das versuche ich, außen vor zu lassen und auf das einzugehen, was der Schüler selbst an Ideen entwickelt hat. Wenn ich meinen Erwartungshorizont habe und ein Schüler trotzdem noch mal in eine andere Richtung schreibt, dass ich auch versuche, das positiv gelten zu lassen als Ersatz für das, was ich eigentlich erwartet hatte.

I Welchen Eindruck hast du von einer Schülerleistung, wenn du unmittelbar davor eine sehr gute oder auch eine sehr schlechte Klausur bewertet hast?

L Also grundsätzlich versuche ich mich davon nicht beeinflussen zu lassen. Da ich den Erwartungshorizont habe, versuche ich relativ sachlich die Punkte zusammen zu zählen und mich davon nicht beeinflussen zu lassen.

I Hast du den Eindruck, dass deine Kollegen ihre Schüler strenger oder milder bewerten als du? Wo und woran machst du das fest?

L Auf jeden Fall gibt es das, dass ich merke, dass Kollegen zum Beispiel die, die die Klasse im letzten Jahr hatten strenger oder milder bewertet haben. Dementsprechend, wenn dann Rückmeldungen kommen wie „Ich hatte bis jetzt immer eine Eins und jetzt plötzlich eine Drei“, also daran würde ich es festmachen, auch an der Rückmeldung der Schüler. Und manchmal daran, ob man merkt, dass sie sehr gut gelernt haben oder auch nicht, weil sie dachten, na ja, es wird bestimmt ganz einfach. Daran würde ich es auch so ein bisschen festmachen, ob ein Kollege vorher eher mild oder streng war.

I Hat deiner Meinung nach das Schriftbild einen Einfluss auf deine Bewertung und wenn ja, welchen?

L Eigentlich soll es nicht. Es kann sein, dass es unterbewusst trotzdem einen Einfluss hat, aber ich versuche, wie gesagt, das außen vor zu lassen und am Erwartungshorizont mich entlang zu hangeln.

I Und wenn ein Schüler dir sympathisch oder unsympathisch ist, welchen Einfluss hat dies auf deine Bewertung?

L Das sollte eigentlich keinen Einfluss haben. Wahrscheinlich hat es trotzdem, leichte Einflüsse. Aber auch hier versuche ich bei jeder Klausur neu zu überlegen oder dem Schüler eine Chance zu geben. Was heißt eine Chance zu geben? Bei Unsympathischen heißt es ja nicht unbedingt, dass sie schlecht sind aber manchmal hängt es damit zusammen, dass ein Schüler, der unmotiviert ist, der sich nicht viel beteiligt manchmal auch unsympathischer ist als jemand, der engagiert mitmacht. Von daher hat es natürlich schon so ein bisschen Einfluss. Aber wie gesagt, durch die Kriterien, die ich vorher aufstelle, habe ich schon eine Möglichkeit einigermaßen objektiv dranzugehen.

I Welche Aufgabenniveaus berücksichtigst du in deiner Klausur?

L Das musst du noch mal kurz erklären.

I Welche Schwierigkeitsgrade?

L Ich versuche den Schwierigkeitsgrad nicht an den Besten anzulegen, sondern grundsätzlich an dem, was leistbar ist. Sozusagen vom Großteil der Klasse. Daran versuch ich die Klausur auszurichten und nicht eine Expertenklausur zu schreiben. Ich

unterteile aber nicht in den einzelnen Aufgaben noch mal nach Schwierigkeit, sondern ich versuche schon für jeden Aufgabentyp eine Schwierigkeit aufzustellen, an der sie sich abarbeiten können, die sie aber eigentlich nicht überfordern sollte aber auch nicht unterfordern.

I Und variierst du dann die Schwierigkeitsgrade oder ist, jede Aufgabe gleich schwer?

L Ja, im Grunde ist jede Aufgabe gleich schwer. Nur, dass natürlich immer die Schüler es leichter finden, Texte zusammenzufassen als noch mal eine eigene Meinung zu sagen und das zu erläutern. Das hängt vom Schüler ab, ob er es schwerer findet oder nicht.

I An welchem Maßstab orientierst du dich bei der Bewertung der Schülerleistung?

L Am Maßstab des Erwartungshorizonts, den ich mir vorher gesetzt habe, und finde in dem Fall auch die Musterlösung.

I Und, was ist, wenn du merkst, die ganze Klasse, es hat keiner eine Eins und keiner Eine Zwei erreicht. Schwächst du das dann ein bisschen ab und sagst: Ok, ich bin dann weicher mit der Punkteverteilung, dann muss man halt nicht fünfzehn Punkte bekommen, um die fünfzehn Punkte zu kriegen sondern vielleicht elf oder zwölf?

L Das kommt darauf an, wenn ich merke, ich habe schlecht formuliert, so wie es jetzt bei der Klausur war. Da habe ich schon gemerkt, die dritte Aufgabe war schlecht formuliert und unklar, wenn ich die mit in die Bewertung reinnehmen würde, würde die Klausur wesentlich schlechter ausfallen. Da wäre dritte Aufgabe an sich bei jedem schlechter gewesen als die eigentliche Leistung, und da habe ich mich dafür entschieden, sie rauszunehmen. Ansonsten kommt es ein bisschen auf die Klasse an. Wenn ich merke, es ist eine sehr engagierte Klasse, die motiviert mitmacht und dann sieht plötzlich die Klausur oder Klassenarbeit schlecht aus, dann muss ich noch mal hinterfragen, ob sie vielleicht doch zu schwer angelegt war und ich das irgendwie angleichen muss. Oder aber, ob es eine Klasse ist, die grundsätzlich auch nicht so gut mitgearbeitet hat, und wo ich denke, das ist so ein bisschen die Konsequenz davon.

I Ok, das heißt, wenn eine Klasse nicht so gut mitgearbeitet hat, dann bist du nicht so geneigt dich erweichen zu lassen und sonst eher runterzugehen. Und wie sieht es aus, wenn du merkst da hat sich ein Schüler echt viel Mühe gegeben aber ihm fehlen noch noch ein oder zwei Punkte. Du würdest ihn schon gerne motivieren wollen, weil er sich echt in den letzten Wochen sehr angestrengt hat, aber es nicht hinhaut, nach deinem Erwartungshorizont? Wie gehst du da vor?

L Im schriftlichen Bereich würde ich es dann trotzdem entsprechend des Erwartungshorizonts durchziehen und das dann eher über die mündliche Leistung berücksichtigen. Dann eher zu schauen, inwieweit kann ich das dann positiv werten.

I Das heißt, du würdest ihm einfach mal eine gute Note in der mündlichen Mitarbeit geben?

L Wenn es gerechtfertigt ist, genau. Wenn ich sehe da bemüht sich jemand, dann, ist das ja eben auch eher die mündliche Bemühung. Wenn ich das in der Klausur merke, auch auch den Unterschied zwischen der letzten und der aktuellen, dass ich da schon eine

Anstrengung sehe, dann spiegelt sich das in der Note wider. Und sei es durch den Ausdruck oder es muss sich an irgendwelchen Kriterien, die im Erwartungshorizont drinstehen widerspiegeln.

Curriculum Vitae

Vorname Name:	Nadine Schlomske
Geburtsdatum, Ort:	24.10.1980 in Freiburg i. Breisgau
Familienstand:	ledig

Schulbildung

2000	Abitur
1997-2000	Walter-Eucken-Wirtschaftsgymnasium
1991-1997	Weiherhof-Realschule
1987-1991	Weiherhof-Grundschule

Studium

12/2007	Abschluss: Magistra Artium
2002 - 2008	Studium der Kognitionswissenschaft
2000 – 2008	Studium der Erziehungswissenschaft und der Anglistik an der Albert-Ludwigs-Universität Freiburg

Beruf

seit 06/2012	wissenschaftliche Mitarbeiterin am Friedl Schöller-Stiftungslehrstuhl für Unterrichts- und Hochschulforschung an der TUM School of Education München
04/2008 - 05/2012	wissenschaftliche Mitarbeiterin am Lehrstuhl für Schulpädagogik und Didaktik an der Friedrich-Schiller-Universität Jena

Ehrenwörtliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Dissertation mit dem Titel: „Technologiegestützte Leistungsdiagnostik in Schule und Hochschule“ selbstständig und ohne unerlaubte Hilfe angefertigt habe. Mir ist die Promotionsordnung der Friedrich-Schiller-Universität Jena bekannt.

Ich versichere, dass ich ausschließlich die angegebenen Quellen und Hilfen in Anspruch genommen habe.

Die Dienste eines Promotionsberaters habe ich nicht in Anspruch genommen.

Bei der Auswahl und Auswertung des Materials sowie in Bezug auf die Herstellung des Manuskripts wurde ich von meinen Betreuern, Frau Prof. Dr. M. Gläser-Zikuda und Herrn PD Dr. Pablo Pirnay-Dummer unterstützt. Darüber hinaus haben keine Dritten unmittelbar und mittelbar geldwerte Leistungen von mir für Arbeiten erhalten, die im Zusammenhang mit der vorliegenden Dissertation stehen.

Ich versichere außerdem, die vorliegende Forschungsarbeit noch nicht als Prüfungsarbeit für eine staatliche - oder wissenschaftliche Prüfung eingereicht zu haben und die des Weiteren weder in ihrer Gesamtheit noch in Teilen oder ähnlicher Form bei einer anderen Hochschule bzw. einer anderen Fakultät als Dissertation eingereicht zu haben.

Ich versichere, die gemachten Angaben nach bestem Wissen gemacht, die Wahrheit gesagt und nichts verschwiegen zu haben.

München, den 05.10.2011

Nadine Schlomske