EFFEKTE ELABORIERTER FEEDBACKS AUF DAS TEXTVERSTEHEN:

Untersuchungen zur Wirksamkeit von Feedbackinhalten unter Berücksichtigung des Präsentationsmodus in computerbasierten Testsettings

Inaugural-Dissertation

in der Fakultät Humanwissenschaften der

Otto-Friedrich-Universität Bamberg

vorgelegt von

Dipl.-Psych. Stefanie Golke

aus

Görlitz

Bamberg, den 17.10.2012

Tag der mündlichen Prüfung: 06.02.2013
Dekan: Universitätsprofessor Dr. Stefan Lautenbacher
Erstgutachterin: Universitätsprofessorin Dr. Cordula Artelt
Zweitgutachterin: Universitätsprofessorin Dr. Barbara Drechsel

Danksagung

An dieser Stelle möchte ich mich ganz herzlich bei den Personen bedanken, die an der Entstehung der vorliegenden Arbeit entscheidend beteiligt waren. In erster Linie möchte ich mich ganz besonders bei meiner Betreuerin Prof. Dr. Cordula Artelt und Herrn Prof. Dr. Tobias Dörfler bedanken, die durch ihre hohe fachliche Kompetenz und ihr Engagement das DFG-Projekt, in dessen Rahmen diese Arbeit entstanden ist, erfolgreich auf die Beine gestellt und mir durch die Übertragung dieses Dissertationsprojektes ihr Vertrauen geschenkt haben. Frau Artelt danke ich für ihre verlässliche und engagierte Unterstützung in allen Phasen der Arbeit. Mit konstruktiver Kritik und präzisem Sachverstand hat sie diese Arbeit maßgeblich angeregt und vorangetrieben. Mindestens genauso in die Betreuung der Arbeit eingebunden war Tobias Dörfler. Ich danke ihm für seine fachlichen Impulse, seine Geduld und sein sprichwörtlich stets offenes Ohr – nicht nur für inhaltliche Belange der Arbeit. Auch wenn es beim Ringen um die Inhalte des Projekts manchmal hitzig wurde, und ich dir, lieber Tobias, dabei das eine oder andere graue Haar beschert haben mag, bleibt mir die Zusammenarbeit als ein sehr kollegiales und freundschaftliches Miteinander in Erinnerung.

Weiterhin möchte ich mich herzlich bei Prof. Dr. Barbara Drechsel für Ihre spontane Bereitschaft bedanken, diese Arbeit zu begutachten und mir mit Rat und Tat zur Seite zu stehen.

Mein besonderer Dank gilt darüber hinaus Peter Kuntner und Gabriel Freise, die die computerbasierte Umgebung für das Herzstück dieser Arbeit, die Experimente, ermöglicht haben. Gabriel Freise hat den Grundstein für das Programm gelegt und mit dem Verlassen des Lehrstuhls seine Arbeiten an Peter Kuntner übergeben, der es durch sein Können und seine Geschicklichkeit maßgeblich und beharrlich in seine letztendliche Form gebracht hat. Er hat stets die Ruhe bewiesen und mich nie im Regen stehen lassen, wenn es kurz vor den Experimenten mal wieder eng wurde und letzte Änderungen nötig waren. Lieber Peter, vielen Dank dafür!

Nicht zuletzt möchte ich mich bei den (ehemaligen) StudentInnen bedanken, die mich in der Vorbereitung und Durchführung der Untersuchungen dieser und darüber hinausgehender Arbeiten unterstützt haben und damit wichtige Stützen für das Gelingen des Projektes waren. Im Laufe der letzten fünf Jahre sind es so viele gewesen, dass ich

ihre Namen hier nicht alle aufzählen kann. Aber stellvertretend für alle möchte ich als langjährige Hilfskräfte Natalie Weber, Heike Hofgräff, Katharina Wiesmann und Ingrid Eder herausheben und ihnen nochmals ausdrücklich meinen Dank für alles, was sie geleistet haben, aussprechen.

Inhaltsverzeichnis

1 Problemstellung und Ziel der Arbeit

Die vorliegende Arbeit ist Teil des DFG-Projektes „Dynamisches Testen im Bereich der Lesekompetenz – Zur Diagnostik und Beeinflussbarkeit der Lesekompetenz durch Feedback und (meta-) kognitive Hilfen in einer computerbasierten Untersuchung“ (AR 301/7- 1, AR 301/7-2), das in das DFG-Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ eingebunden war. Dynamische Tests verfolgen das Ziel, aufbauend auf der Idee von Vygotsky (1964), neben dem aktuellen Entwicklungs- bzw. Leistungsstand auch das Entwicklungs- oder Leistungspotential eines Testanden zu erfassen. Dazu werden in den Test Leistungsrückmeldungen bzw. Hilfestellungen implementiert. Die Reaktionen des Testanden auf diese Maßnahmen werden genutzt, um das individuelle Leistungspotential abzuschätzen (Sternberg & Grigorenko, 2002).
Dynamische Tests sind über ihren typischen Anwendungsbereich der Intelligenzdiagnostik hinaus in vielen Leistungsbereichen einsetzbar. Spezifischer Ausbaubedarf besteht im Bereich der Lesekompetenz und dafür ist unter anderem die Identifikation anwendbarer und effektiver Rückmeldungen (Feedbacks) unter den spezifischen Rahmenbedingungen des Tests notwendig (Dörfler, Golke & Artelt, 2009). Hieran knüpft die vorliegende Arbeit an. Ihr Ziel besteht also darin, die Effekte von Feedbacks auf das Textverstehen/die Lesekompetenz unter spezifischen Rahmenbedingungen experimentell zu untersuchen.

Der Begriff Feedback steht für ein breites Spektrum an Leistungsrückmeldungen und Hilfestellungen. Sein Wirkungsprinzip basiert auf der Plastizität und Adaptivität von Lernprozessen. Dabei wird das wesentliche Potential in der Korrektur von Fehlern gesehen (Kulhavy & Stock, 1989). Feedback zu bereits gelingenden Prozessen (z.B. richtig beantwortete Aufgaben) ist weniger erkenntnisreich und produktiv. Das Überwinden eines falschen Verständnisses oder defizitärer Prozesse bedeutet Lernfortschritt.
In der Regel wird davon ausgegangen, dass ausführlichere, spezifische Hinweise, so genannte elaborierte Feedbacks, besser geeignet sind, um Lernen und Leistung zu fördern, als einfache Rückmeldungen, die beispielsweise nur mitteilen, dass ein Fehler

gemacht wurde. Das gilt insbesondere für kognitiv anspruchsvolle Fähigkeitsbereiche wie Textverstehen.

Feedbackstudien im Bereich der Lesekompetenz, die sich tatsächlich auf die Förderung der vor allem hierarchiehöheren Prozesse des Textverstehens beziehen, sind rar, haben aber einige effektive Rückmeldungen hervorgebracht. Diese basieren typischerweise auf mehrmaligen Interventionssitzungen mit zusätzlichen Strategietrainings und die Feedbacks sind dann entsprechend auch meist auf die Nutzung der (meta-)kognitiven Strategien angepasst. Diese Merkmale sind mit den spezifischen Anforderungen des Einsatzes von Feedback in einem Test, mit dem Ziel auch die Reaktionen auf die Hilfestellungen zu erfassen (Dynamischer Test), allerdings nicht vereinbar.

Die benötigten Feedbacks müssen vor allem kurz und prägnant sein und oft – pro Item – gegeben werden, damit in der Testung möglichst viele Antworten, das heißt Datenpunkte, vom Lerner gesammelt werden können. Außerdem kann in der Testung kein Dialog stattfinden und der Einbezug eines Trainings ist kaum möglich. Gleichzeitig muss es gelingen, Rückmeldungen zu konstruieren, die trotz dem beschränkten Interventionsspielraum bestmöglich die defizitären Prozesse (Feedback setzt an Fehlern an) fördern. Angewendet auf die Anforderungen des Textverstehens ergeben sich weitere Herausforderungen.

Textverstehen ist ein hoch komplexer Prozess der Bedeutungskonstruktion. Verstehen bedeutet zu erfassen, was gemeint ist. Dafür muss der Leser aktiv und unter Nutzung seines Vor- bzw. Weltwissens Sinnzusammenhänge des Textes konstruieren. Der Aufbau einer Bedeutungsvorstellung ist ein ressourcenfordernder Informationsverarbeitungsprozess, der zudem maßgeblich von den Rezeptionszielen und den Einstellungen des Lesers abhängt (Zwaan & Rapp, 2006).

Textverstehen ist von zentraler Bedeutung, in schulischen wie außerschulischen Situationen. Aber viele Leser, von der Grundschule bis ins Erwachsenenalter, haben Schwierigkeiten ein umfassendes und tieferes Verständnis über einen gelesenen Text aufzubauen. Eine Quelle von Verständnisschwierigkeiten liegt in defizitären Prozessen der Textverarbeitung (Allington & McGill-Franzen, 2009; Best, Rowe, Ozuru & McNamara, 2005).

Das Ansetzen an diesen Verständnisschwierigkeiten mittels Feedback ist möglich. Die besonderen Herausforderungen für diese Arbeit ergeben sich aus den spezifischen, notwendigen Rahmenbedingungen der Feedbackintervention.

Zur Untersuchung der Effekte von Feedbacks auf das Textverstehen wurden zwei aufeinander aufbauende Experimente mit Schülern der sechsten Klassenstufe durchgeführt. Der Fokus des ersten Experiments liegt auf den Inhalten der Feedbacks, im zweiten Experiment wird auch der Präsentationsmodus berücksichtigt.

Überblick über die Arbeit

Im theoretischen Teil der Arbeit werden zunächst die Grundlagen des Textverstehens skizziert. Im Anschluss daran werden ausführlicher Theorien und Ansätze der Feedbackforschung dargelegt; ein wichtiger Bestandteil ist die Darstellung der Feedbackstudien im Bereich des Textverstehens. Vor dem Hintergrund der Theorien und Befunde beider Forschungsbereiche und unter Berücksichtigung der Rahmenbedingungen, die sich in Hinblick auf den geplanten Dynamischen Test ergeben, werden schließlich potentiell effektive Feedbackinterventionen für das Textverstehen abgeleitet und beschrieben.

Der empirische Teil der Arbeit gliedert sich in die Darstellung beider Experimente, die in sich geschlossen beschrieben und diskutiert werden. Die Arbeit endet mit der abschließenden Diskussion beider Experimente und einem Ausblick auf weiterführende Forschungsfragen und die Bedeutung der gewonnen Erkenntnisse für die Entwicklung des Dynamischen Tests der Lesekompetenz.

2 Textverstehen

Das Lesen und das Verstehen eines Textes sind nicht gleichbedeutend. Lesen bezieht sich zunächst „nur“ auf die Fähigkeit, Grapheme in Phoneme umsetzen zu können (Lesefertigkeit). Verstehen im Rahmen der Textverarbeitung bedeutet aber, zu erfassen, was im Text gemeint ist. Der Leser muss aktiv unter Rückgriff auf sein Vor- bzw. Weltwissen eine Vorstellung von der Bedeutung der gelesenen Informationen aufbauen. Verstehen ist also ein stark konstruktiver Prozess. „Der Text ist (…) nicht Träger von Bedeutungen“ (Schnotz, 2006, S. 224) und das Verstehen stellt sich nicht, wie vom deterministischen Kommunikationsmodell in den Anfängen der Sprachverarbeitungsforschung postuliert (vgl. Christmann & Groeben, 1999), vollständig durch das Dekodieren der schriftlichen Informationen ein. Stattdessen dient der Text

> vielmehr als Auslöser für mentale Konstruktionsprozesse, die teils von der externen Textinformation und teils von der internen (im semantischen Gedächtnis gespeicherten) Vorwissensinformation angeleitet werden. Diese Konstruktionsprozesse führen zum Aufbau einer mentalen Repräsentation des im Text beschriebenen Sachverhalts, was subjektiv als ‚Erfassen der Textbedeutung‘ erlebt wird (Schnotz, 2006, S. 224).

Das Konstruieren von Textbedeutung ist ein komplexer Prozess, der sich aus mehreren, miteinander interagierenden und dabei flexiblen Teilprozessen zusammensetzt, die vom Identifizieren der Wortbedeutungen über das Verstehen eines Satzes bis hin zum Erfassen der Bedeutung größerer Textteile bzw. des Textes als Ganzes reichen. Die Ausführung der Teilprozesse mündet im Normalfall im Aufbau mehrerer mentaler Repräsentationen, die verschiedentlich die Inhalte bzw. die Bedeutung des Textes abbilden. Häufig wird vom Leser ein möglichst umfassendes, tiefes Verständnis eines Textes erwartet. Dies setzt komplexe Informationsverarbeitungsprozesse voraus, ist aber auch von den Zielsetzungen und Erwartungen des Lesers abhängig.

Textverstehen ist also gekennzeichnet durch kognitive Konstruktivität, Komplexität und Flexibilität, aber auch durch eine gewisse Variabilität seiner „Ergebnisse“. Das Lesen und Verarbeiten von Texten kann auch bei adäquatem Rezeptionsziel des Lesers zu einem unzureichenden, fehlerhaften oder gänzlich ausbleibenden Verständnis führen. Aufgrund der Komplexität des Textverstehens können die Quellen von Verständnisschwierigkeiten

vielfältiger Art sein. Mit Blick auf den Gegenstandsbereich dieser Arbeit sind Verständnisschwierigkeiten, insbesondere jene auf der Textebene, auch als die Anknüpfungspunkte für die Feedbackinterventionen von Interesse.

Im Folgenden werden zunächst die Prozesse und Ebenen der Bedeutungskonstruktion erörtert. Im Anschluss daran werden mögliche Quellen von Verständnisschwierigkeiten erläutert und daran anknüpfend Ansatzpunkte für Interventionen skizziert.

2.1 Prozesse der Textverarbeitung

In den Theorien und Ansätzen der Textverarbeitung wird der Leseprozess typischerweise in mehrere Teilprozesse aufgegliedert und diese werden wiederum der Wort-, Satz- oder Textebene zugeordnet (Christmann & Groeben, 1999). Der Leseprozess beginnt auf der Wortebene mit der Identifikation von Buchstaben und Wörtern sowie der Erfassung der Wortbedeutungen.
Das Erkennen von Buchstaben und Wörtern läuft im Wesentlichen über die visuelle Verarbeitung. Bezüglich der Art der Verarbeitung (vgl. Balota, Yap & Cortese, 2006 für Positionen/Ansätze) wird inzwischen weitestgehend übereinstimmend davon ausgegangen, dass *abstrakte* Buchstabeneinheiten eher parallel statt seriell verarbeitet werden. Wörter werden, zumindest sofern sie bereits bekannt, das heißt im Gedächtnis gespeichert sind, über den direkten visuellen Zugang via Aktivationsausbreitung (interaktives Aktivationsmodell von McClelland & Rumelhart, 1981) identifiziert (Christmann & Groeben, 1999; Richter & Christmann, 2002).
Das Identifizieren und Erfassen der Wortbedeutungen wird bereits deutlich durch seinen Kontext gelenkt. Ein Wort ist Teil eines Satzes, der einen bestimmten semantischen Gehalt vermittelt, und dieser Kontext beeinflusst das Erfassen der Bedeutung des einzelnen Wortes (Carpenter, Miyake & Just, 1995; Sanford, 2002).

Das Verstehen eines Satzes (Satzebene) erfordert, über die Prozesse der Wortebene hinaus, Wortfolgen aufeinander zu beziehen und dabei in ein strukturiertes Gefüge zu bringen (Christmann & Groeben, 1999). Dazu werden die Wortfolgen primär über die Analyse ihrer semantischen, aber auch ihrer syntaktischen Relationen (vgl. Irmen, 2006 für Theorien und Analysestrategien) zu Propositionen integriert.

Der Begriff der Proposition bezieht sich in kognitionspsychologischen Sprachverarbeitungstheorien auf hypothetische, komplexe Symbole (Schnotz, 2006), die kognitive, keine sprachlichen Bedeutungseinheiten reflektieren (Christmann & Groeben, 1999; vgl. auch Kintsch, 1974; van Dijk & Kintsch, 1983). „A proposition refers to a state, event, or action and may have a truth value with respect to a real or imagery world." (Graesser, Millis & Zwaan, 1997, S. 168) Eine Proposition besteht aus einem Relationssymbol, dem so genannten Prädikat, und aus einem oder mehreren Symbolen für Entitäten (Gegenstände oder Ereignisse), den so genannten Argumenten. Verschiedene Merkmale der Textoberfläche wie Zeitform, Artikel oder Passiv-/Aktiv-Konstruktionen werden nicht über die Proposition repräsentiert (Kintsch & van Dijk, 1978; van Dijk & Kintsch, 1983).
Die Relevanz der Propositionen im Rahmen der Textverarbeitung gilt als belegt (van Dijk & Kintsch, 1983). Das Satzverstehen besteht also in der Bildung von Propostionen aus der zugrundeliegenden Satzstruktur, wobei nur die Bedeutung (vorerst) erhalten bleibt, die Form, das heißt die Textoberfläche, verschwindet dagegen (Christmann & Groeben, 1999).

Das Verstehen eines Textes (Textebene) geht wiederum über das Erfassen der Bedeutungen einzelner Sätze hinaus. Der Sinngehalt eines Textes ergibt sich nicht automatisch aus der Summe der einzelnen, verarbeiteten Sätze, sondern deren Bedeutungsinhalte müssen zusätzlich aufeinander bezogen, integriert und dabei in einen sinnvollen, kohärenten Zusammenhang gebracht werden. Kohärenz ergibt sich dem Leser, wenn er die im Text beschriebenen Ereignisse, Handlungen und Zustände miteinander in Verbindung bringen kann (Singer, Graesser & Trabasso, 1994). Bezüglich des Herstellens von Kohärenz werden zwei Aspekte unterschieden: die lokale und die globale Kohärenzbildung.
Lokale Kohärenz wird hergestellt, indem aufeinanderfolgende Propositionen bzw. Sätze miteinander verknüpft werden. Meistens wird die lokale Kohärenzbildung auf die Integration der eingehenden Information mit den ein bis drei vorherigen Sätzen bezogen. Sie kann sich aber ebenso auf das Verbinden einer eingehenden mit einer im Arbeitsgedächtnis gehaltenen Information beziehen (Cook, Halleran & O'Brien, 1998; Graesser et al., 1997), ohne Rücksicht auf die Nähe der Informationen in der Textoberfläche.

Bei längeren und komplexeren Texten wird neben der lokalen auch die globale Kohärenzbildung erforderlich (Richter & Christmann, 2002). Um den globalen Zusammenhang des Textes bzw. seiner Abschnitte auf höherer Abstraktionsebene erfassen zu können, muss der Leser Propositionssequenzen weiter verdichten und wiederum miteinander verknüpfen. Durch die Anwendung der so genannten Makroregeln –Auslassen bzw. Selegieren sowie Generalisieren und Konstruieren – werden während des Lesens Gruppen von Mikropropostionen zu Makropropositionen verdichtet, das Resultat wird als Makrostruktur bezeichnet. Aufgrund der Rekursivität der Makroregeln kann die über die Makrostruktur repräsentierte Textbedeutung unterschiedlich abstrakt aufgebaut werden (Kintsch & van Dijk, 1978; van Dijk, 1980; van Dijk & Kintsch, 1983). Der Aufbau der Makrostruktur, also die Reduktion des Textes, ist ein stark konstruktiver Prozess und er erfolgt auf der Grundlage des Textes sowie unter Beteiligung text-, vorwissens-, interessens- und zielbasierter Inferenzen (Christmann & Schreier, 2003). Die Bedeutung der Makropropositionen für das Textverstehen zeigt sich darin, dass diese besser im Gedächtnis verhaftet bleiben als die ihnen zugrunde liegenden Mikropropositionen (Kintsch, 1998).
Der Aufbau einer Makrostruktur kann durch die Kenntnis und Wahrnehmung so genannter Superstrukturen unterstützt werden (Kintsch & van Dijk, 1978). „Superstrukturen beschreiben im Sinne eines Rasters oder abstrakten Schemas (…) die globale Ordnung von Texten, die eine spezifische, konventionalisierte Struktur haben“ (Richter & Christmann, 2002, S. 33), wie sie zum Beispiel bei wissenschaftlichen Aufsätzen vorzufinden ist. Sie sind als Regeln oder Kategorien mental gespeichert und steuern als Erwartungen den Leseprozess im Sinne einer vorwissensgeleiteten Verarbeitung (Kintsch & van Dijk, 1978; Richter & Christmann, 2002). Des Weiteren können für den Aufbau einer angemessenen Textbedeutung auch das Erkennen und Verstehen rhetorischer, stilistischer und argumentativer Strategien, die vom Autor meist zur Akzentuierung bestimmter Inhaltselemente eingesetzt werden, relevant sein (Richter & Christmann, 2002).

Das Herstellen von Kohärenz kann automatisch oder strategisch sowie wissens- oder textbasiert erfolgen (Richter & Christmann, 2002). Textbasierte Hinweise über die Relation von Informationen bieten syntaktische, semantische und konzeptuelle Mittel (Christmann & Groeben, 1999). Ein Beispiel einer textbasierten Relationsart ist die Koreferenz, die vorliegt, wenn in aufeinanderfolgenden Sätzen auf denselben

Referenten Bezug genommen wird. Koreferenz kann durch verschiedene semantische oder syntaktische Mittel erzeugt werden, beispielsweise die Rekurrenz (Wortwiederholung), die pronominale Koreferenz (Pronomen für Nomen), Anapher oder Katapher (Rück- bzw. Vorverweis). Eine globale Strategie der Koreferenz ist die Thema-Rhema- bzw. *Given-New*-Strategie, die besagt, dass eine neue Information (Rhema bzw. New-Anteil) auf eine bereits gespeicherte Information (Thema bzw. Given-Anteil) zu beziehen ist (Graesser et al., 2007; Graesser, McNamara & Louwerse, 2003; Zwaan & Singer, 2003). Die semantische Relation zwischen Sätzen oder Textabschnitten kann des Weiteren durch Mittel konzeptuell-inhaltlicher Art angezeigt werden, beispielsweise über die Konzeptualisierung als Konzept-Beispiel oder These-Antithese (Sanders, Spooren & Noordman, 1992; Schnotz, 2006).
Je mehr eindeutige Hinweise zum Zusammenhang zwischen Sätzen bzw. Passagen in einem Text enthalten sind, desto besser gelingt dem Leser im Allgemeinen der Aufbau einer satzübergreifenden Bedeutungsvorstellung. An den Stellen eines Textes, an denen keine entsprechenden Signale enthalten sind, obliegt es dem Leser unter Rückgriff auf sein Vor- bzw. Weltwissen entsprechende Schlussfolgerungsprozesse (Inferenzen) anzustellen (Graesser, Singer & Trabasso, 1994).

Textverstehen ist untrennbar mit dem Ziehen von Inferenzen verbunden (Graesser et al., 2003). Inferenzielle Prozesse sind auf hierarchieniedriger als auch hierarchiehoher Ebene von besonderer Bedeutung, um eine kohärente Textrepräsentation aufbauen zu können (Schnotz, 2006). Unter Inferenzen werden jene Verstehensprozesse gefasst, die über die explizite sprachliche Information hinausgehen, indem sie zusätzliche, verstehensrelevante Propositionen anreichern, die gegebene Textinformation strukturieren oder verdichten. Dabei spielen das Sprach- und Weltwissen des Lesers sowie der situative Kontext der aktuellen Textinformation eine große Rolle (Blanc & Tapiero, 2001; Graesser et al., 1994; Halldorson & Singer, 2002; Kintsch, 1988; Noordman & Vonk, 1992).
In der Literatur werden mehrere Arten von Inferenzen (auf der Grundlage verschiedener Kategorisierungen) unterschieden (Graesser et al., 1997; Graesser et al., 1994; Singer, 1994; van den Broek, 1994). Wesentlich für den Aufbau einer kohärenten Textrepräsentation sind die so genannten Brücken- oder kohärenzstiftenden Inferenzen, denn sie verbinden die aktuelle Textinformation mit dem vorherigen Text. Dabei wird entweder „eine Brücke“ zwischen einem Nomen oder Pronomen und dem vorherigen

Text „geschlagen“ oder eine kausale Beziehung zwischen zwei Sätzen hergestellt (Long, Seely, Oppy & Golding, 1996; Singer, Harkness & Stewart, 1997). Es ist davon auszugehen, dass Brückeninferenzen routinemäßig während des Lesens (*online*) gezogen werden (Trabasso & Magliano, 1996; Trabasso, Suh, Payton & Jain, 1995).

Neben den Brückeninferenzen sind auch die elaborativen Inferenzen von Bedeutung für das (tiefere) Textverständnis. Elaborative Inferenzen sind jene Prozesse, die ausgehend von expliziter Textinformation zusätzliche Informationen aus dem Langzeitgedächtnis anreichern. Dazu werden beispielsweise gezählt: Konsequenzen von Ereignissen und Handlungen, Eigenschaften von Objekten, Mittel/Instrumente zur Ausführung von Handlungen, räumliche Beziehungen zwischen Objekten/Agenten und Motive/Pläne für Handlungen (Graesser et al., 1997; Graesser et al., 1994). Das Bilden dieser Inferenzen wird zwar durch den konkreten Kontext angeregt, aber sie sind für das Herstellen oder Aufrechterhalten der Kohärenz nicht erforderlich. Im Gegensatz zu Brückeninferenzen treten elaborative Inferenzen nicht zwangsläufig während der Textverarbeitung auf (Casteel, 1993; Garrod, O'Brien, Morris & Rayner, 1990; Graesser et al., 2007; O'Brien, E. J., Shank, Myers & Rayner, 1988).

In der Frage, welche Inferenzen online oder *offline* bzw. unter welchen Bedingungen gezogen werden, werden auf der Grundlage empirischer Befunde verschiedene Positionen vertreten (Graesser & Zwaan, 1995; Long et al., 1996; Singer et al., 1994). Meistens werden in der Literatur die minimalistische Hypothese (McKoon & Ratcliff, 1992) und die konstruktivistische Theorie (Graesser et al., 1997; Graesser et al., 1994) gegenübergestellt und diskutiert. Nach der minimalistischen Hypothese (McKoon & Ratcliff, 1992) werden nur jene Inferenzen online generiert, „[that are] based on easily available information and those required for local coherence“ (McKoon & Ratcliff, 1992, S. 441). Darüber hinausgehende Inferenzen werden der Hypothese nach dagegen nur gezogen, wenn der Leser ein entsprechendes Rezeptionsziel verfolgt oder spezifische Lesestrategien einsetzt. Demgegenüber postuliert die konstruktivistische Theorie (Graesser et al., 1997; Graesser et al., 1994), dass beim Lesen jene Inferenzen gebildet werden, die a) Kohärenz auf lokaler sowie globaler Ebene herstellen (Albrecht & O'Brien, 1993; Long, Graesser & Golding, 1992), b) Erklärungen für im Text beschriebene Ereignisse, Handlungen oder Zustände schaffen (Trabasso & Magliano, 1996; Trabasso & Wiley, 2005) und c) den Rezeptionszielen des Lesers dienen (vgl. Graesser et al., 1994; Singer, 1994).

Diese Diskussion tritt jedoch dann in den Hintergrund, wenn das Textverstehen mithilfe von konkreten Fragestellungen überprüft wird. Wenn ein Leser eine Fragestellung präsentiert bekommt, die nach einer bestimmten Inferenz fragt (z.B. nach einer Ursache für ein Ereignis oder nach dem Motiv einer Handlung), wird dadurch auch das Inferieren dieser Information angeregt. Inferenzen, die durch eine Aufgabenstellung bzw. durch Nachfragen induziert werden, „provide insight both into the differential availability of information from the text and into the cognitive processes that can operate on this information" (van den Broek, 1994, S. 557). Das heißt, auf diesem Weg wird erfasst, ob bestimmte Sinnzusammenhänge hergestellt werden können, nicht, ob sie vom Leser selbstständig für den Aufbau eines (tieferen) Textverständnisses generiert würden (Singer, 1994).

2.2 Ebenen der Bedeutungskonstruktion

In den kognitionspsychologischen Ansätzen des Textverstehens wird davon ausgegangen, dass der Rezipient während und nach dem Lesen verschiedene mentale Repräsentationen aufbaut (Graesser et al., 1997; Kintsch, 1998; van Dijk & Kintsch, 1983; Zwaan, 1996). Für gewöhnlich werden dabei drei Repräsentationsebenen unterschieden: die Oberflächenebene (*surface level*), die propositionale Repräsentationsebene (*text base*) und die Modellebene (*situation model* oder *mental model*). Deren Unterscheidbarkeit ist durch eine Reihe von Untersuchungen untermauert (z.B. Kintsch, Welsch, Schmalhofer & Zimny, 1990; Schmalhofer & Glavanov, 1986; Zwaan, 1994). Darüber hinaus werden mitunter auch zwei weitere, von Graesser und Kollegen (1997) vorgeschlagene Ebenen berücksichtigt, die so genannte Kommunikationsebene (*communication level*) und die Genreebene (*text genre level*). Die Kommunikationsebene bezieht sich auf die Merkmale, die sich aus dem kommunikativen Kontext eines Textes ergeben. Die Genreebene beinhaltet die strukturellen Merkmale, die aus der Art des Textes resultieren und für die Bedeutungskonstruktion von Relevanz sind.

Alle drei (bzw. fünf) Repräsentationsebenen sind für das Textverstehen relevant und tragen in einem komplexen Zusammenspiel zur Bedeutungsbildung und -repräsentation während und nach der Lesephase bei (Graesser et al., 1997). Sie unterscheiden sich in ihren Repräsentationseigenschaften und Funktionen (Schnotz,

2006). Die nachfolgenden Erläuterungen beschränken sich auf die drei etablierten, empirisch gut fundierten Ebenen: die Oberflächenrepräsentation, die propositionale Repräsentation und das Situationsmodell.

Oberflächenrepräsentation

Auf der Oberflächenebene wird die Textoberfläche, also der exakte Wortlaut und die Syntax des gelesenen Textabschnitts, abgebildet. Sie ermöglicht das wortwörtliche Wiederholen von Textteilen, auch ohne ein Verständnis des Gelesenen erreicht haben zu müssen (Schnotz, 2006). Die Oberflächenrepräsentation bezieht sich im Normalfall auf den zuletzt gelesenen Satz (Graesser et al., 1997) und wird nur für kurze Zeit (Kintsch & Bates, 1977; Kintsch et al., 1990), schätzungsweise ein paar Sekunden lang (Zwaan & Singer, 2003), im Arbeitsgedächtnis behalten.

Unter bestimmten Umständen kann die Oberflächenrepräsentation aber auch länger und umfangreicher im Gedächtnis behalten werden (vgl. auch Kintsch & Bates, 1977), etwa wenn die Textoberfläche von hoher Relevanz für die Bedeutung bzw. das Verständnis eines Satzes/Textabschnitts oder von einer bestimmten pragmatischen Relevanz (z.B. Witze, Beleidigungen, textbezogene Erwartungen) ist (Long, 1994; Murphy, G. L. & Shapiro, 1994; Zwaan, 1994). Die erhöhte Erinnerungsleistung liegt dann darin begründet, dass der Leser seine Aufmerksamkeit verstärkt auf die Textoberfläche richtet, wodurch die Informationen besser enkodiert und damit im Gedächtnis behalten und erinnert werden können (vgl. auch Kintsch & Bates, 1977).

Die Oberflächenrepräsentation gilt als grundlegend für die weitere Verarbeitung des Gelesenen im Sinne der Bedeutungskonstruktion. Das Erfassen der Bedeutung eines Textes ist maßgeblich mit dem Aufbau der Textbasis und insbesondere des Situationsmodells verbunden.

Propositionale Repräsentationsebene

Die propositionale Repräsentationsebene bzw. die Textbasis (eingeführt von Kintsch, 1974) beinhaltet den semantischen Gehalt eines Textes bzw. Textabschnitts in Form einer hierarchisch strukturierten Menge von Propositionen (vgl. Abschnitt 2.1), ergänzt um einige wenige hierarchieniedrige Inferenzen, die für die Herstellung der Kohärenz notwendig sind (van Dijk & Kintsch, 1983).

Wird nur die Textbasis konstruiert, hat der Leser erfasst, was im Text gesagt wird, ohne dabei zu verstehen, welcher Sachverhalt damit gemeint ist. Dieser Fall kann am ehesten bei sehr schweren Texten oder bei der Beschreibung unmöglicher Sachverhalte auftreten (Schnotz, 2006). Schnotz führt hier zur Illustrierung ein Beispiel von Chomsky (o.J.) an: „Farblose grüne Ideen schlafen wütend." (S. 226) Dieser scheinbar sinnlose Satz lässt sich im Sinne der propositionalen Textverarbeitung in Propositionen zerlegen und es lassen sich im Zusammenhang mit dem Satz einige herkömmliche Fragen beantworten (Was wird getan? Wer schläft? etc.). Das Ungewöhnliche ist jedoch, dass der in dem Satz beschriebene Sachverhalt nicht vorstellbar ist, es existiert dafür keine Relation im semantischen Gedächtnis (Schnotz, 2006). Ähnlich verhält es sich bei sehr schweren Texten. Hier kann keine Vorstellung über die Aussagen im Text aufgebaut werden, weil dem Leser die dafür notwendigen Verknüpfungen der Textinformationen mit seinem Wissen nicht gelingen. Das kann wiederum darin begründet sein, dass dem Leser das entsprechende Wissen fehlt, also keine Referenten vorhanden sind, und/oder die Textstruktur Verknüpfungen zu vorhandenem, relevantem Wissen erschwert oder unterbindet (Stichwort: Textverständlichkeit).

Im Rahmen der propositionalen Textverarbeitung wird davon ausgegangen, dass vor allem die zuletzt verarbeiteten und wichtigsten Propositionen im Arbeitsgedächtnis behalten werden (Kintsch & van Dijk, 1978); entsprechend wird diese Repräsentationsform auch als *gist-like memory* (Kintsch & van Dijk, 1978) bezeichnet. Die propositionale Repräsentation ist weniger flüchtig als die Oberflächenrepräsentation. Dennoch wird sie nicht sehr lang behalten, was sich darin äußert, dass die Leser bald nach der Textrezeption zunehmend weniger gut unterscheiden können, welche Aussagen so im Text (Textbasis) enthalten sind und welche Aussagen zwar im Sinne des Textes richtig, aber inferiert sind (Radvansky, 2005).

Modellebene

Wie bereits angedeutet, wird zusätzlich zur Textbasis noch eine weitere Repräsentationsform angenommen, die die konstruierte Vorstellung über das, was im Text gemeint ist, reflektiert. Es gibt eine Reihe kognitiver Phänomene, die nur mit einer propositionalen Repräsentation nicht hinreichend gut erklärbar sind (vgl. van Dijk & Kintsch, 1983, S. 338-342; auszugsweise: Zwaan & Radvansky, 1998, S. 163-

165). Es ist nicht die Repräsentation der Textbasis allein, sondern die unter Rückgriff auf das Vor- bzw. Weltwissen *konstruierte* Repräsentation (Bedeutungsvorstellung) der im Text dargestellten Situation(en), die das Verstehen eines Textes ausmachen (Kintsch, 1994; Zwaan & Radvansky, 1998). Diese Ebene der mentalen Repräsentation ist mit den Begriffen Situationsmodell (van Dijk & Kintsch, 1983) und mentales Modell (Johnson-Laird, 1983) belegt; beide werden weitestgehend synonym verwendet.
Zur Illustrierung der Bedeutung der Modellebene in Abgrenzung zur Textbasis wird ein Beispiel aus Sanford und Garrod (1998, zitiert nach Zwaan & Singer, 2003, S. 92f.) herangezogen. Gegenstand dieser Studie ist der Vergleich zwischen Satzpaaren, die in ihrer propositionalen Struktur ähnlich oder gleich sind, durch den Austausch einzelner Wörter aber einen grundsätzlich verschiedenen Sachverhalt übermitteln:

„Harry put the wallpaper on the table. Then he put his mug of coffee on the paper."

Diese beiden Sätze lassen sich problemlos integrieren. Beim Lesen dieser wurde sehr wahrscheinlich eine (hier vor allem räumliche) Vorstellung aufgebaut, in der die Tapete (wallpaper) auf dem Tisch (table) liegt und der Kaffeebecher (mug of coffee) auf der Tapete abgestellt ist. Diesem Satzpaar wurde in der Untersuchung ein weiteres gegenübergestellt:

„Harry put the wallpaper on the wall. Then he put his mug of coffee on the paper."

In diesem zweiten Satzpaar ist gegenüber dem ersten Paar nur ein Wort verändert (wall/Wand anstatt table/Tisch). Die propositionale Struktur beider Satzpaare ist vergleichbar, doch es werden unterschiedliche Sachverhalte transportiert. Der Punkt ist, dass – obwohl auch das zweite Satzpaar auf propositionaler Ebene gut integrierbar ist – die meisten Leser im Gegensatz zum ersten Satzpaar stocken: ein Gegenstand, hier der Kaffeebecher, kann nicht auf eine vertikale Fläche (die Tapete ist *an* der Wand angebracht) gestellt werden. Die dargestellte Situation ist nicht mit dem Hintergrundwissen der Leser vereinbar. Basierte das Verstehen von schriftlicher Information allein auf dem Verknüpfen von Propositionen, würden keine Schwierigkeiten beim Lesen und Verstehen auftreten.
Das Vorgehen bei Sanford und Garrod ist typisch für den Forschungsbereich; neben Satzpaaren werden bei vergleichbarem Vorgehen auch kürzere Texte eingesetzt. Diese

und ähnliche Untersuchungen haben dazu beigetragen, dass die postulierte Ebene des Situationsmodells bzw. mentalen Modells letztlich anerkannt wurde. Auch wenn es bislang keinen „absolut zwingenden Beweis für die Existenz mentaler Modelle gibt“ (Schnotz, 2006, S. 227), ist die Vielzahl der vorliegenden Befunde zum Textverstehen durch die zusätzliche Annahme der Modellebene schlüssiger zu erklären (Schnotz, 2006; van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998).

Bezüglich der Struktur oder Art der mentalen Repräsentation der Modellebene differieren die Annahmen (Zwaan & Singer, 2003). Im Wesentlichen lassen sich die Positionen darauf verkürzen, ob die Repräsentation als eher analoger oder symbolischer Art gesehen wird. Das Situationsmodell nach van Dijk und Kintsch ist als propositionale und damit symbolische Repräsentation konzeptualisiert; innere Bilder sind hier auf der Modellebene möglich (Kintsch, 1998; van Dijk & Kintsch, 1983). Das mentale Modell nach Johnson-Laird ist als nicht-propositionale, sondern analoge Repräsentation beschrieben (Johnson-Laird, 1989) – „mental models […] have a structure that corresponds to the perceived or conceived structure of [a] state of affairs” (Johnson-Laird, Herrmann & Chaffin, 1984, S. 311). In ähnlicher Weise fasst Barsalou (1999) die Modellebene als eine analoge Repräsentation auf. Beim Lesen werden die im Text enthaltenen Situationen/Gegebenheiten vom Leser wahrnehmungsgetreu mental simuliert und im Fortlauf des Textes entsprechend abgeändert – „as in a perception of a physical scene“ (Barsalou, 1999, S. 605).

Hinsichtlich der Inhalte von Situationsmodellen besteht weitestgehend Konsens darüber, dass Rezipienten beim verstehenden Lesen (wenigstens) die folgenden fünf Dimensionen der im Text angesprochenen Situationen routinemäßig und/oder strategisch (Graesser et al., 1997) überwachen: Raum, Zeit, Entitäten (Agenten bzw. Objekte), Kausalität und Motivation (Zwaan & Radvansky, 1998; Zwaan & Singer, 2003). Die Dimensionen weisen deutlich Abhängigkeiten untereinander auf. Beispielsweise geht ein Ortswechsel in einer Geschichte häufig auch mit dem zeitlichen Fortschreiten des Geschehens einher und Ursachen von Ereignissen hängen davon ab, was genau (durch wen) passiert.

Es hat sich gezeigt, dass beim Lesen mehrere der Dimensionen simultan überwacht werden (Zwaan, 1998; Zwaan, Magliano & Graesser, 1995). Über Interaktionen der verschiedenen situationalen Dimensionen und deren Auswirkungen auf das Verständnis ist noch sehr wenig bekannt (Therriault & Rinck, 2007); die bisherige Forschung zu den inhaltlichen Aspekten von Situationsmodellen ist weitestgehend eindimensional ausgerichtet (Zwaan & Radvansky, 1998).
Den Dimensionen kommt beim Lesen sehr wahrscheinlich eine unterschiedliche Stellung zu (Therriault & Rinck, 2007). Auf Nachfrage oder bei entsprechender Aufgabenstellung können zwar alle Dimensionen gleich gut beachtet werden, soweit die Fähigkeiten und das Wissen des Lesers es ihm ermöglichen. Aber beim Lesen ohne Leseanlass bezüglich eines konkreten Aspekts finden sich Unterschiede hinsichtlich der „Bevorzugung" bestimmter Aspekte/Dimensionen. So steht die räumliche Dimension bei der Situationsmodellbildung normalerweise am wenigsten im Vordergrund (Zwaan, 1998). Kausale und motivationale Aspekte scheinen dagegen am ehesten überwacht zu werden (Zwaan, 1996). Beim Lesen von Erzähltexten übernehmen Leser sehr wahrscheinlich die Perspektive des Protagonisten und behalten bevorzugt die für den Protagonisten relevanten Informationen (Morrow & Bower, 1989; O'Brien, E.J. & Albrecht, 1992). Für die Relevanz der Überwachung der Protagonisten und der Zeit spricht eine neuere Untersuchung von Therriault, Rinck und Zwaan (2006), die zeigte, dass beide Dimensionen auch dann überwacht werden, wenn die Leser auf die Beachtung anderer Aspekte instruiert waren.

Die Bildung der Situationsmodelle hat grundsätzlich Prozesscharakter. Dies impliziert, dass die Repräsentationen kumulativ entstehen, aber nicht nur durch die fortwährende Anhäufung von Informationen, sondern auch durch Reorganisation oder Korrektur des bestehenden Modells infolge neuer Informationen (Kintsch, 2009).
Eine konkretere Idee vom Prozess der Situationsmodellbildung bietet das *general processing framework* von Zwaan und Radvansky (1998). Unterschieden wird hier zwischen drei Arten des Situationsmodells: *current model*, *integrated model* und *complete model*. Beim Lesen eines ersten Satzes oder Satzteils eines vorliegenden Textes entsteht mental repräsentiert eine Vorstellung, ein Modell, vom soeben Gelesenen bezüglich (eines) Agenten, des Raums und/oder einer der anderen situationalen Dimensionen. Dieses Modell ist das aktuelle Modell (current model). Es bezieht sich

immer auf die gerade rezipierte Information (normalerweise ein Satz oder Satzteil), es ist "the model currently under construction" (Zwaan & Radvansky, 1998, S. 165) und hat daher keinen Bestand. Mit dem Rezipieren des nächsten (Teil-)Satzes entsteht für diesen ein neues aktuelles Modell. Sein Inhalt wird anhand bestimmter Kriterien mit dem ehemals aktuellen Modell des vorherigen (Teil-)Satzes verknüpft, wodurch das integrierte Modell (integrated model) entsteht. Mit jeder neuen aufgenommenen Information wird das integrierte Modell aktualisiert. Das ist ein fortlaufender Prozess. Die neuen Informationen werden in die bestehenden des integrierten Modells eingegliedert, indem Verbindungen zwischen dem aktuellen Modell und relevanten Aspekten des integrierten Modells hinsichtlich der verschiedenen Dimensionen geschaffen werden.

Dabei werden aber nicht alle Informationen repräsentiert bzw. nicht alle eingehenden Informationen werden genutzt und in das bestehende Modell integriert. Ausschlaggebend ist hierbei die vom Leser vor dem Hintergrund des bestehenden, integrierten Modells wahrgenommene Relevanz von Informationen. Auf der Basis von linguistischen Hinweisreizen im Text oder dem Weltwissen des Lesers, das bedeutet seinem Wissen über erfahrene/erlebte Situationen hinsichtlich beispielsweise Kausalitäten oder Motivationen für bestimmte Handlungen, werden passende, relevante Informationen in den Vordergrund gestellt. Zwaan und Radvansky benutzen hierfür das Konzept des *foregrounding*: „information is foregrounded by creating and maintaining a retrieval cue to this information in [short term working memory]" (S. 167). Die so im Kurzzeitgedächtnis aktiv gehaltene Information ist im Situationsmodell zugänglicher und andere, nicht relevante bzw. unpassende Informationen treten in den Hintergrund. Wenn keine Verbindung zwischen der neuen Information und der zuvor beschriebenen (bereits repräsentierten) Situation ersichtlich ist, wird das Modell nicht aktualisiert bzw. die neue Information wird nicht integriert. Dabei vermuten Zwaan und Radvansky auch, dass für die verschiedenen situationalen Dimensionen auch unterschiedliche Kriterien der Relevanz gelten: für die Dimension Zeit ist die zeitliche Nähe zwischen Ereignissen entscheidend, für die Dimension der Motivation dagegen beispielsweise sehr häufig ein noch nicht vollendetes Ziel oder eine nicht abgeschlossene Handlung eines Agenten.

Am Ende eines Textes, wenn alle beinhalteten Informationen gelesen und verarbeitet sind, wird das dann bestehende integrierte Modell zum vollständigen Modell (complete model), das aber nicht zwangsläufig auch das finale Modell ist. Denn durch Reflektieren

über den Text oder nochmaliges Lesen des Textes können durch den Leser weitere Aktualisierungen am Situationsmodell vorgenommen werden.

Welche und wie viele Informationen letztendlich jedoch auf der Modellebene mental repräsentiert und behalten werden, ist äußerst variabel. Situationsmodelle können von abstrakten, groben Skizzen einer Situation einerseits bis hin zu lebensechten Modellen andererseits reichen (Graesser & Zwaan, 1995). Perfetti und Britt (1995) weisen in diesem Zusammenhang daraufhin, dass Situationsmodelle nicht nur von ihren Inhalten, sondern auch von deren Dynamik (*focus*, *contrastiveness* und *definiteness*) bestimmt werden, und meinen damit die Genauigkeit, die „Kontrastschärfe“ und den Fokus des Modells.
Bedingt wird die Detailliertheit bzw. Abstraktheit eines Modells durch die Interaktion zwischen den expliziten Textmerkmalen, der Aufgabe (Leseanlass) und den Fähigkeiten und Kapazitäten des Lesers, insbesondere seinem Vor- und Weltwissen, aber auch seinen Rezeptionszielen (Guthrie & Wigfield, 2000). Grundsätzlich gilt, dass das Textverständnis umso tiefer ausfällt, je mehr Bezüge zwischen dem Text und dem Wissen hergestellt werden (Graesser et al., 1994). Doch das setzt stark ressourcenfordernde Verarbeitungsprozesse voraus, die unter anderem schon den normalen Lesefluss deutlich verlangsamen (Rinck, 2000; Zwaan & Radvansky, 1998).
Es ist davon auszugehen, dass im Normalfall keine detailreichen, die realen Gegebenheiten widerspiegelnde mentalen Repräsentationen gebildet werden, auch wenn dies dem Leser prinzipiell möglich ist. Leser tendieren im Allgemeinen eher zu einem „ökonomischen“ Rezeptionsstil (Rinck, 2000) – Situationsmodelle sind vermutlich nur so detailliert und präzise, wie die Rezeptionsziele es erfordern (Foertsch & Gernsbacher, 1994). Sparsame Repräsentationen der im Text beschriebenen Sachverhalte sind daher wohl eher die Regel.

Im Vergleich zu den mentalen Repräsentationen der Textoberfläche und der Textbasis wird das Situationsmodell deutlich länger behalten und dominiert dadurch den Textverarbeitungsprozess (Radvansky, 2005). Der Aufbau eines Situationsmodells bedeutet, dass ein Verständnis des Gelesenen zumindest in Teilen erlangt werden konnte.
Das gebildete Modell ist aber nicht nur ein Endprodukt des Rezeptionsprozesses, sondern es unterstützt – im Leseverlauf – seinerseits auch das Verstehen der weiteren schriftlichen

Informationen. „Inferences can be made in the process of constructing a situation model, and situation models can influence the nature of the inferences that will be made." (Zwaan & Radvansky, 1998, S. 163) Auf der Modellebene ist die globale Kohärenz des Textes besser zu überwachen als auf der Ebene der propositionalen Repräsentation (Tapiero & Otero, 2002). Das Situationsmodell unterstützt das Verstehen auch dadurch, dass es den Kontext für eingehende Informationen bietet; implizit ist diese Eigenschaft schon in den vorangegangenen Ausführungen zur Konstruktion und speziell den Aktualisierungen von Situationsmodellen enthalten. Dabei beeinflusst das bereits konstruierte Verständnis in bestimmter Weise den Fokus der Aufmerksamkeit bei der weiteren Textverarbeitung (Rinck, 2000; Zwaan & Radvansky, 1998).

Anhand des gebildeten und gespeicherten Situationsmodells können wiederum Informationen abgerufen werden, beispielsweise um auf vorgegebene (Test-)Fragen zu antworten. Dabei erleichtert die Struktur des Modells den Abruf (Tapiero & Otero, 2002). Die Bildung des Situationsmodells ermöglicht nicht nur das Wiedergeben expliziter und impliziter Aussagen des Textes, sondern auf seiner Grundlage können die Aussagen eines Textes auch interpretiert und bewertet werden (Rinck, 2008).

2.3 Quellen von Verständnisschwierigkeiten

Das Vorhandensein von Verständnisschwierigkeiten beim bzw. nach dem Lesen eines Textes bedeutet, dass der Leser eine (in Teilen) inkohärente Bedeutungsrepräsentation auf der Ebene der Textbasis und/oder des Situationsmodells aufgebaut hat. Die Konstruktion einer Bedeutungsvorstellung ist, wie in den beiden vorangegangenen Kapiteln bereits erläutert wurde, kein Automatismus. Leser tendieren zu einer in Hinsicht auf ihre Rezeptionsziele bzw. Aufgabenstellungen angepasste, ökonomische Verarbeitung von Texten. Dabei entstehen oft eher „sparsame" Situationsmodelle, das heißt, es werden nur wenige Inferenzen gezogen und es wird kein (tieferes) Verständnis bezüglich der im Text beschriebenen Sachverhalte konstruiert (vgl. Abschnitt 2.2). Weniger kohärente und/oder elaborierte mentale Repräsentationen reflektieren also nicht zwangsläufig Verständnisschwierigkeiten. Bei entsprechender Instruktion oder auf Nachfragen können die verschiedenen Inferenzen beispielsweise oft generiert werden (Graesser et al., 1994; Singer & Leon, 2007). Nichtsdestotrotz belegen unzählige Studien, dass nicht wenige Leser (massive) Schwierigkeiten darin aufweisen, (trotz entsprechender

Absicht) ein hinreichend gutes Textverständnis zu erlangen (Allington & McGill-Franzen, 2009; Artelt, Stanat, Schneider & Schiefele, 2001; Artelt, Stanat, Schneider, Schiefele & Lehmann, 2004; Cain & Oakhill, 2007b; Nation, 2005). Differenzielle Unterschiede im Textverstehen/der Lesekompetenz sind in verschiedenen Altersgruppen zu finden, von Grundschülern über adoleszente und bis hin zu erwachsenen Lesern (Allington & McGill-Franzen, 2009).

Verständnisschwierigkeiten können prinzipiell in allen Faktoren begründet sein, die in ihrem Zusammenspiel auch die erfolgreiche Bildung des Situationsmodells bedingen: Merkmale des Textes, des Lesekontextes und/oder des Lesers (van den Broek & Kremer, 1999). Van den Broek und Kremer fassen zusammen, dass seitens des Textes sowohl sein Inhalt als auch seine Struktur das Verstehen für den Leser behindern können (vgl. auch Christmann, 1989). Bleibt ein Text in seinen Ausführungen vage oder wirkt aufgrund geringer oder fehlender Kohäsion (Graesser et al., 2007) verwirrend, wird der Leser im Allgemeinen kaum Anknüpfungspunkte für die Relationen zu seinem Vorwissen extrahieren können. Der Bedeutungsgehalt des Textes wird ihm nicht oder nur schwer erfassbar sein. Ein ähnliches Problem stellt sich dem Leser, wenn er einen Text mit für ihn unbekannten Konzepten liest, wie es primär bei Sachtexten anstatt Erzähltexten vorkommen kann (Best et al., 2005; Graesser et al., 2003). Das Hintergrundwissen des Lesers bietet in diesem Fall nur (sehr) wenige Anknüpfungspunkte für die Inhalte im Text.
In Bezug auf die vorliegende Arbeit wird der Bereich der Textmerkmale als Quelle für Verständnisschwierigkeiten aber weitestgehend ausgeschlossen. Alle Texte sind im Vorfeld der Untersuchung hinsichtlich ihrer Verständlichkeit und Angemessenheit für die Stichprobe der Sechstklässler überprüft worden. Gleichzeitig wurden nur Texte eingesetzt, die für die Altersgruppe kein spezifisches Vorwissen erfordern (vgl. Abschnitt 5.4).

Zu den Merkmalen des Lesekontexts, die dem Aufbau eines Textverständnisses abträglich sein können, zählen van den Broek und Kremer (1999) unwichtige und deshalb potentiell ablenkende Aufgaben sowie vor allem missverständliche oder nicht explizit gemachte Erwartungen bzw. Aufgabenstellungen an den Leser. Damit ist gemeint, dass der Leser irrtümlich mit einem anderen als dem vom Lehrer oder Testleiter gemeinten Rezeptionsziel einen Text liest und bearbeitet und infolgedessen das

Ergebnis des Leseprozesses mit den Erwartungen von außen sehr wahrscheinlich nicht übereinstimmt. Dabei handelt es sich offenkundig um ein generelles Problem der Instruktion, nicht nur bezogen auf Schwierigkeiten beim Textverstehen. Ungleiche Erwartungen werden im Rahmen von Verständnisschwierigkeiten auf Textebene, auf die sich van den Broek und Kremer beziehen, dann relevant, wenn der Leser ein oberflächigeres Rezeptionsziel als das an ihn gestellte einnimmt. In diesem Fall müssen dem mangelnden Textverständnis also keine defizitären Fähigkeiten des Lesers zugrunde liegen.
Der Problematik ungleicher Rezeptionsziele ist augenscheinlich durch eindeutige, explizite Instruktionen entgegenzuwirken. Außerdem ist auf die Passung zwischen einer Aufgabenstellung (z.B. Testfrage) und den anvisierten Teilprozessen zu achten (van den Broek & Kremer, 1999) – interessiert ein tiefes Textverständnis, muss die entsprechende Aufgabenstellung auch dieses ansprechen und nicht etwa die Angabe einer wortwörtlichen Textinformationen. Da diese Aspekte des Lesekontextes in der vorliegenden Arbeit berücksichtigt sind (vgl. Abschnitte 5.4 und 5.6), wird das Risiko missverständlicher oder unpassender Instruktionen bzw. Fragestellungen als Quellen von Verständnisschwierigkeiten als vernachlässigbar angesehen.

Wesentlich für die vorliegende Arbeit sind dagegen Merkmale des Lesers, die ursächlich für Verständnisschwierigkeiten sein können. Hierzu zählen das Vor- bzw. Weltwissen des Lesers, seine allgemeinen kognitiven Fähigkeiten bzw. Ressourcen (z.B. Arbeitsgedächtniskapazität, Aufmerksamkeit) und insbesondere seine Fähigkeiten im Rahmen der Textverarbeitung. Der Faktor unzureichenden Vor- bzw. Weltwissens wird aus bereits genanntem Grund in den weiteren Betrachtungen vernachlässigt. Allgemeine kognitive Fähigkeiten bzw. Ressourcen wie die Arbeitsgedächtniskapazität beeinflussen zweifelsohne das Textverstehen (Artelt, Stanat et al., 2001; Just & Carpenter, 1992; Yuill, Oakhill & Parkin, 1989). Sie sind jedoch keine Anknüpfungspunkte der Interventionen dieser Arbeit. Ihren Einflüssen auf das Textverstehen wurde allerdings in der Umsetzung der Interventionen soweit wie möglich Rechnung getragen (z.B. Sichtbarkeit der Texte und des Feedbacks bei Antwortkorrekturen; vgl. Abschnitte 5.4-5.6).
Der Fokus der Interventionen in dieser Untersuchung liegt auf den Verständnisschwierigkeiten, die in den Fähigkeiten bzw. Prozessen der Textverarbeitung selbst liegen. Ihnen wird im Rahmen der Forschung zu

Verständnisschwierigkeiten beim Lesen einerseits (vgl. Cain & Oakhill, 2007a; van den Broek & Kremer, 1999) und Interventionsmaßnahmen andererseits (vgl. Artelt et al., 2005; Cain & Oakhill, 2007a; Paris, Wasik & Turner, 1991) große Aufmerksamkeit gewidmet. Ob Verständnisschwierigkeiten ursächlich auf basale, hierarchieniedrige Prozesse zurückzuführen sind (Perfetti, 1994; Perfetti, Marron & Foltz, 1996) oder doch (auch) auf hierarchiehoher Prozessebene begründet sind (Cain & Oakhill, 1999; Oakhill & Yuill, 1996), wird kontrovers diskutiert. Perfetti und Kollegen (1996) argumentieren, dass Probleme bei der Konstruktion eines umfassenderen Textverständnisses aus defizitären, und zwar fehlerhaften und unzureichend automatisierten, Worterkennungsprozessen resultieren. Für die basalen Prozesse auf Wortebene benötigen die betreffenden Leser entsprechend mehr kognitive Ressourcen, die dann nicht mehr für hierarchiehöhere Prozesse zur Verfügung stehen. In der Tat finden sich auch Korrelationen zwischen Worterkennungsprozessen und dem Textverstehen (Artelt, Schiefele & Schneider, 2001). Auf der anderen Seite weisen nicht alle Leser mit Verständnisschwierigkeiten auch defizitäre hierarchieniedrige Prozesse auf (Oakhill, 1993; Oakhill & Yuill, 1996; van den Broek & Kremer, 1999), das heißt, trotz routinisierter Worterkennungs- und gegebenenfalls auch Integrationsprozesse auf Satzebene können auf der Textebene Verständnisschwierigkeiten auftreten (Allington & McGill-Franzen, 2009; Nation, 2005).

Verständnisschwierigkeiten, die auf der hierarchiehöheren Prozessebene verankert sind, werden sehr häufig in Zusammenhang mit unzureichenden inferenziellen Prozessen gebracht (Nation, 2005). Es liegen mehrere Studien vor, die belegen, dass schlechtere Leser weniger Inferenzen ziehen als gute Leser (Cain, 1999; Cain, Oakhill, Barnes & Bryant, 2001; Long, Oppy & Seely, 1997; Oakhill, 1982) und es gibt Hinweise, wonach schwache Leser zudem irrelevante Informationen weniger gut unterdrücken können (Gernsbacher & Faust, 1991; Gernsbacher, Varner & Faust, 1990).

Die Schwierigkeiten schwacher Leser beim Herstellen von Inferenzen erstrecken sich sowohl auf textbasierte Inferenzen zum Herstellen lokaler Kohärenz als auch auf wissensbasierte Inferenzen zum Erschließen impliziter Informationen im Text (Cain & Oakhill, 1999; Oakhill, 1984). Ihnen gelingt es also weniger gut, eine kohärente, integrierte Repräsentation eines Textes zu konstruieren. Auf der Oberflächenebene unterscheiden sie sich dagegen nicht von besseren Lesern (Oakhill, 1984). Als mögliche Ursachen werden die folgenden Faktoren diskutiert: eine geringere

Arbeitsgedächtniskapazität, fehlendes Hintergrundwissen und das Auslassen der notwendigen konstruktiven Prozesse des Verbindens und Integrierens von text- und wissensbasierten Informationen.

Die Befunde sprechen dafür, dass eine geringere Arbeitsgedächtniskapazität keine hinreichende Erklärung für die Defizite darstellt. In Oakhill (1984) konnte auch das Vorlegen des Textes die Probleme beim Herstellen der geforderten Inferenzen nicht beheben. In Cain und Oakhill (1999) zeigte sich dagegen, dass sich die vergleichsweise geringe Inferenzleistung der schwachen Leser mit der Vorlage des Textes verbesserte (vgl. auch Ozuru, Best, Bell, Witherspoon & McNamara, 2007) – hinsichtlich der Testaufgaben zu textbasierten Inferenzen sogar bis auf das Niveau der guten Leser (allerdings liegt hier möglicherweise ein Deckeneffekt in der Untersuchung vor). Aber bei den wissensbasierten Inferenzen blieben die schwachen Leser schlechter als die Vergleichsgruppe, hier brachte also auch die Möglichkeit des Nachlesens im Text keinen wesentlichen Vorteil. Wurden die Leser dagegen explizit auf die für die Aufgabe relevanten Textstellen hingewiesen, verbesserten sie sich auch hier, blieben aber immer noch unter der Leistung der guten Leser.

Dass den Lesern das Herstellen der Verknüpfungen nach dem Aufzeigen der notwendigen Informationen im Text besser gelang, spricht nach Cain und Oakhill (1999) für die generelle Fähigkeit der Leser zum Generieren auch von elaborierten Inferenzen. Stattdessen vermuten die Autorinnen, dass die schwachen Leser beim Lesen der Texte eine Lesemotivation aufbauen und/oder Strategien einsetzen, die für ein tieferes Verständnis des gelesenen Textes nicht tragend sind. Dieser Aspekt überschneidet sich mit der bereits erläuterten Problematik inadäquater Rezeptionsziele im Rahmen der Kontextmerkmale. Beispielsweise könnten sich Leser auf ein wortgenaues, flüssiges Lesen konzentrieren anstatt das Verstehen während des Lesens zu überwachen. Ebenso konnte bezüglich der wissensbasierten Inferenzen gezeigt werden, dass die schwachen Leser bis zu einem gewissen Maß auch in der Lage waren, die für die Inferenz jeweils notwendige Informationen aufzuzeigen, und sie besaßen nachweislich auch das entsprechende Hintergrundwissen. Da sie dann dennoch nicht in der Lage sind, die entsprechenden Inferenzen ohne Hilfe zu generieren, schlussfolgern Cain und Oakhill, dass schwache Leser eher nicht wissen, wann und wie sie ihr Vorwissen mit dem Text verknüpfen sollen.

Weitere Studien stützen die Annahme, dass eine Schwäche schlechter Leser darin besteht, die für die Inferenz notwendigen Informationen zu erkennen (Cain et al., 2001) und/oder die Verknüpfungen zwischen Textinformationen und/oder Referenten im Langzeitgedächtnis herzustellen (vgl. Morris & Bransford, 1982). Vereinzelte Untersuchungen kommen zwar zu dem gegenteiligen Ergebnis, dass die Verständnisschwierigkeiten nicht auf Defizite beim Integrieren von Informationen zurückgehen (Spooner, Gathercole & Baddeley, 2006). Insgesamt sprechen die Befunde jedoch im Sinne von Cain und Oakhill (1999) dafür, dass Verständnisschwierigkeiten mit dem hinsichtlich Menge und Routinisierungsgrad unzureichenden Generieren von kohärenzstiftenden und elaborativen Inferenzen einhergehen (Coté, Goldman & Saul, 1998; Graesser et al., 2003; Oakhill & Cain, 2007; Oakhill & Garnham, 1988).

Neben den Inferenzen wird auch die metakognitive Überwachung beim Lesen als wichtiger Faktor im Zusammenhang mit Verständnisschwierigkeiten gesehen (Nation, 2005; Oakhill & Cain, 2007). Guten Lesern gelingt die metakognitive Überwachung besser als schwachen Lesern (Cain & Oakhill, 2007b). Perfetti und Kollegen (Perfetti et al., 1996) argumentieren in dem Zusammenhang wiederum, dass die schwachen Leser, die nur ein unzureichendes oder kein Verständnis bezüglich eines Textes aufbauen können, „lediglich“ falschen Annahmen über die Rezeptionsziele aufsitzen; ihnen ist also gegebenenfalls nicht bewusst, dass es darum geht, die Bedeutung des Textes zu erfassen. Möglicherweise konzentriert sich dieses Problem aber auf jüngere Leser (Myers & Paris, 1978). Des Weiteren diskutieren Perfetti und Mitarbeiter (1996) vor dem Hintergrund von Perfettis Theorie der verbalen Effizienz (Perfetti, 1985), dass schwache Leser eventuell (auch) nicht in der Lage sind, die Gesamtheit der Informationen, die sie verarbeiten, mental zu repräsentieren. Allerdings unterstreicht die Mehrheit der Befunde die Bedeutung defizitärer Prozesse als Quellen von Verständnisschwierigkeiten und dabei sind auf der Ebene hierarchiehöherer Prozesse, wie bereits erläutert, die Verstehensüberwachung und insbesondere das Herstellen von text- und wissensbasierten Inferenzen hervorzuheben.

2.4 Ansatzpunkte für Interventionen bei Verständnisschwierigkeiten

Bezugnehmend auf die Bedeutung defizitärer Fähigkeiten des Lesers bei der Textverarbeitung als Ursachenfaktor für Verständnisschwierigkeiten wird an dieser Stelle ein Überblick über jene Ansatzpunkte für Interventionen gegeben, die eben diesen Faktor fokussieren. Die Förderung von Prozessen der Textverarbeitung ist in verschiedenen Interventionen bzw. Trainings eingebunden, die sich grob danach unterscheiden lassen, ob sie einzelne Strategien bzw. Prozesse fokussieren oder mehrere im Verbund einsetzen (vgl. Pearson & Fielding, 1991; Raphael, George, Weber & Nies, 2009).

Im Review des US-amerikanischen *National Reading Panel* (2000) sind sieben Gruppen effektiver Strategien bzw. Techniken zur Förderung des Textverstehens zusammengefasst: Techniken der Verstehensüberwachung, das Beantworten von verständnisprüfenden Fragen, das selbstständige Generieren von Fragen, Strategien zur Inhaltsorganisation, Zusammenfassungen für Texte erstellen sowie Geschichtengrammatiken anwenden und kooperatives Lernen. Dabei handelt es sich um Methoden, die auf unterschiedlichen Wegen im Kern die aktive Auseinandersetzung mit dem Text fordern und fördern (vgl. Pressley, 2000). In diesem Sinn lässt sich auch die Expertise von Artelt und Kollegen (2005) zu Förderansätzen der Lesekompetenz in gekürzter Weise wiedergeben: erfolgreiche Trainings und Instruktionsmaßnahmen in diesem Bereich nutzen Strategien der aktiven Wissensnutzung und der Inhaltsorganisation beim Textverstehen, Strategien des Lokalisierens relevanter Information und des Herstellens von Inferenzen sowie Methoden der Verstehensüberwachung. Als konkrete Methode kann *Self-Explaining* (Chi, 1996; Chi, Leeuw, Chiu & Lavancher, 1994) hervorgehoben werden, das im Prinzip eine Kombination aus Fragen generieren, Vorwissen aktivieren und das Suchen nach Antworten auf die selbstgestellten Fragen darstellt. Diese Methode ist häufig eingebettet in Settings des *Tutorings* und/oder kann durch Anregungen und Hilfestellungen von außen (*Scaffolds/Prompts*) gestützt werden (vgl. auch Abschnitt 3.5.1.1).

Die Interventionen sind also nah an den Prozessen des Textverstehens ausgerichtet. Allerdings fällt dabei auch auf, dass die meisten Interventionen inzwischen auf die Vermittlung und Einübung mehrerer, konzertierter Strategien setzen (z.B. King, 2007; McNamara, Ozuru, Best & O'Reilly, 1995; Raphael & Wonnascott, 1985). Zudem sind

die Fördermaßnahmen oft längerfristiger angelegt (Mateos & Alonso, 1991), um so unter anderem vielfältige Erfahrungen bezüglich der Wirkung und ausreichende Übungs- und Habitualisierungsphasen gewährleisten zu können (Pressley, Borkowski & Schneider, 1989). Interventionen mit einer Dauer von bis zu 12 Stunden gelten dabei als verhältnismäßig kurze Maßnahmen, die unter bestimmten Bedingungen jedoch auch wirksam sein können (Souvignier & Antoniou, 2007).

Die vorliegende Arbeit bewegt sich allerdings in einem Kontext, der ein zusätzliches Training bzw. eine über mehrere Stunden angelegte Interventionsmaßnahme ausschließt (vgl. auch Abschnitt 3.5.3). Stattdessen werden einzelne Hilfestellungen benötigt, die im Rahmen einer Testung beim Lesen von Texten und dem Beantworten entsprechender Verständnisfragen angeboten werden können. Damit ist die Brücke zum Bereich der Feedbackinterventionen zu schlagen.

3 Feedback im pädagogisch-psychologischen Kontext

Der Erwerb und die Anpassung von Wissen, Fähigkeiten und Fertigkeiten bedürfen in weiten Teilen des Austausches zwischen dem Lerner und seiner Umwelt. Rückkopplungsprozessen wird damit in den meisten Theorien des Lehrens und Lernens eine zentrale Stellung eingeräumt (Bangert-Drowns, Kulik, Kulik & Morgan, 1991). Dementsprechend zählt das Geben von Feedback (Rückmeldungen) zu den verbreitetesten psychologischen Interventionen (Kluger & DeNisi, 1998).

In der pädagogisch-psychologischen Literatur finden sich vielzählige Gestaltungsvarianten von Feedback. Entsprechend breit ist das Forschungsfeld, dessen bisheriger Schwerpunkt auf dem Feedback selbst liegt, weniger auf dem Lerner, der Feedback rezipiert. In erster Linie wird sich der Frage gewidmet, welche Art von Feedback am wirksamsten ist bzw. wie Feedback gestaltet sein muss, damit es lern- bzw. leistungsförderlich ist.

Dabei wird intuitiv oft vorausgesetzt, dass Feedback in jedem Fall eine positive, leistungssteigernde Wirkung nach sich zieht. Doch Feedback hat im Allgemeinen sehr variable Effekte auf die Leistung und kann auch leistungshinderlich sein (Kluger & DeNisi, 1996). Latham und Locke (1991), die eine sehr kritische Haltung gegenüber Feedback einnehmen, formulieren:

> Few concepts in psychology have been written about more uncritically and incorrectly than that of feedback. In organizational settings the aphorism 'what gets measured gets done' describes cogently the positive halo surrounding feedback. Actually, feedback is only information, that is, data, and as such has no necessary consequences at all. (S. 224)

Allgemeingültige Aussagen darüber, welche Faktoren sich begünstigend oder hinderlich auf das Lernen bzw. die Leistung auswirken, können nach wie vor nicht getroffen werden. Aus diesem Grund sind eine differenzierte Betrachtung der Befunde der Feedbackliteratur und die angemessene Anwendung des Konzepts von großer Bedeutung. Die Grundlage dafür stellt eine exakte Begriffsklärung dar.

3.1 Begriffsbestimmung

Im pädagogisch-psychologischen Kontext steht der Begriff Feedback für das Prinzip, einem Lerner Rückmeldung über bestimmte Aspekte seiner Leistung oder seines Verhaltens zu geben. Typischerweise bezieht sich Feedback dabei auf die Bearbeitung einer Aufgabenstellung (Hattie & Timperley, 2007; Mory, 2004) und wird daher im Englischen häufig auch als *post-response information* beschrieben (z.B. in Kluger & DeNisi, 1998; Mory, 2004). Als Aufgabenstellung gelten sowohl eng umgrenzte Anforderungen wie beispielsweise eine einzelne Frage zu einem gelesenen Text als auch umfangreiche Aufgaben wie das Verfassen eines Aufsatzes. In jedem Fall kann Feedback im Sinne einer post-response Information gegeben werden; die Art der Rückmeldung würde aber sicherlich unterschiedlich ausfallen.

Jede Feedbackintervention schließt einen Sender, also die Feedbackquelle, und einen Adressaten, an den die Mitteilung gerichtet ist, ein (Ilgen, Fisher & Susan, 1979). Mögliche Feedbackquellen sind nicht nur Personen wie Lehrer, Testleiter oder Peers, sondern auch Lern- oder Testumgebungen, über die Rückmeldungen vermittelt werden (Hattie & Timperley, 2007). Die Intervention kann sowohl an einzelne Personen als auch an mehrere gleichzeitig (z.B. ein Lernduo, eine Schulklasse) adressiert sein. Die meisten Untersuchungen sind jedoch derart konzipiert, dass Lerner individuell angesprochen werden.

Rückmeldungen werden im Normalfall mit der Absicht gegeben, das Lernen zu unterstützen bzw. die Leistung zu fördern (Shute, 2008). Dabei kann darauf abgezielt werden, eine gezeigte Leistung und damit das Verständnis bzw. das Wissen des Lerners zu bestätigen oder zu modifizieren (Mory, 2004). Grundsätzlich kann in einer feedbackunterstützten Lern- und Leistungssituation beides gemeinsam verfolgt werden. Doch in Abhängigkeit von den zugrunde gelegten theoretischen Annahmen darüber, welche Wirkung Feedback im Lernprozess ausübt, ist im Verlauf der Zeit eine unterschiedliche Schwerpunktsetzung in der Feedbackforschung festzustellen (vgl. Kulhavy, 1977; Kulhavy & Stock, 1989). Am Anfang lag der Fokus vor dem Hintergrund des vorherrschenden behavioristischen (operanten) Paradigmas darauf, das zu bestätigen, was ein Lerner richtig gemacht hat, um so Lernprozesse zu unterstützen. Später verschob sich die Aufmerksamkeit auf das kognitivistische Paradigma und wurde

Feedback primär gegeben, um fehlerhafte oder unzureichende Leistungen zu modifizieren (Kulhavy & Stock, 1989; Mory, 2004). Da das Verständnis und die Umsetzung des Feedbackkonzepts wesentlich von seiner Einordnung in die lerntheoretischen Paradigmen geprägt wird, werden die jeweiligen Annahmen über die Wirkung von Feedback auf Lernen und Leistung und die entsprechenden Implikationen für seine Gestaltung und Anwendung in (forschungsrelevanten) Lernsituationen nachfolgend erläutert.

Aus der operanten Perspektive wird Feedback als Verstärker aufgefasst (Kulhavy & Stock, 1989). Es kann dabei prinzipiell sowohl als positive Konsequenz (z.B. Bestätigung, Lob) zum Aufbau als auch als negative Konsequenz (z.B. Tadel) zum Abbau oder zur Unterdrückung von Verhalten eingesetzt werden. Doch vor allem basierend auf Thorndikes (1932) Effektgesetz liegt der Schwerpunkt bei Feedbackinterventionen im Rahmen des operanten Paradigmas auf der Verstärkung erwünschter Verhaltensweisen. In Lern- und Leistungskontexten sind das typischerweise die richtigen Antworten, für die also positives, bestätigendes Feedback gegeben wird (Kulhavy & Wager, 1993). Unerwünschtes Verhalten in Form von Fehlern wird dagegen nicht beachtet (Kulhavy & Stock, 1989).

Dieses behavioristisch geprägte Verständnis von Feedback war die theoretische Grundlage der Feedbackforschung der ersten Zeit (etwa von 1911 bis Ende der 1970er Jahre, siehe Mory, 2004). Allerdings erbrachten die entsprechend konzeptualisierten Studien unsystematische Effekte auf die Leistung und ließen nicht den Schluss zu, dass Feedback als (positiver) Verstärker fungiert (Anderson, Kulhavy & Andre, 1971; Kulhavy, 1977). Ein weiteres Problem wird im Nachhinein darin gesehen, dass Feedbackinterventionen im pädagogisch-psychologischen Kontext die strengen Bedingungen des operanten Wirkprinzips verletzen. So befinden sich Lernende in einer Feedbackintervention nicht unter vergleichbar „potenten Kontingenzen wie physische Deprivation [z.B. Hunger]“ (Kulhavy, 1977, S. 213, *Übersetzg. v. Verf.*), die eine nachfolgende positive Konsequenz zu einem Verstärker werden lassen. Darüber hinaus unterliegen die Stimuli (d.h. Aufgabenstellungen) und Reaktionen (d.h. Antwortverhalten) in Lehr-Lernsituationen ständiger Veränderung, was im strengen Sinn ebenfalls nicht mit einem operanten Setting vereinbar ist (Kulhavy, 1977; Kulhavy & Stock, 1989).

Infolge der Unzulänglichkeiten des operanten Feedbackansatzes wurde das Augenmerk verstärkt auf die vermittelnden kognitiven Prozesse des Lerners gerichtet. Getragen vom

Zeitgeist, der kognitiven Wende, rückte das kognitivistische Paradigma in den Vordergrund. Feedback wird hierbei als Information verstanden, die in die (meta-) kognitiven Verarbeitungsprozesse des Lerners einfließt (Mory, 2004) und durch Überschreiben, Ergänzen oder Restrukturieren von Gedächtnisinhalten zur Veränderung von Wissensstrukturen beiträgt (Butler & Winne, 1995). Rückmeldungen dienen dabei vorrangig der Korrektur von Fehlern und werden folglich primär bei falschen Antworten eingesetzt (Kulhavy & Stock, 1989). Feedback nach richtigen Antworten dient dagegen lediglich der Bestätigung (Kulhavy & Wager, 1993), wodurch es kaum zum Lernfortschritt beiträgt.

Neben dem Schwerpunkt der Fehlerkorrektur betont der kognitivistische Ansatz die Rolle des Lerners, der Feedback nutzen und verarbeiten *kann* (Kulhavy & Wager, 1993). Eine sozusagen automatische Wirkung von Rückmeldungen wie im operanten Modell wird hier nicht angenommen (Latham & Locke, 1991).

Eine stärkere Betonung der Rolle des Lerners im Lernprozess und beim Verarbeiten und Nutzen von Feedback findet im Rahmen des konstruktivistischen Paradigmas statt (Mory, 2004). Der Konstruktivismus geht davon aus, dass es keine objektive Realität, kein objektives Wissen gibt, was vom Lerner aufgenommen würde, sondern Lerner ihre eigene Realität bzw. ihr Wissen konstruieren. Verstehen und Lernen bedeuten, eingehende Informationen auf der Grundlage des eigenen Vorwissens (d.h. Erfahrungen, mentale Strukturen und Überzeugungen) zu interpretieren bzw. zu konstruieren (Jonassen, 1991). In Ableitung daraus wird Feedback als Informations*angebot* verstanden, das vom Lerner individuell wahrgenommen und vor dem eigenen Erfahrungshintergrund verarbeitet wird (Mory, 2004). Dabei muss das Wissen, das in Feedbackinterventionen übermittelt wird, nicht notwendigerweise mit dem Wissen, das der Lerner konstruiert, übereinstimmen (Duffy & Jonassen, 1992; Jonassen, 1991).

Der konstruktivistische Feedbackansatz überlappt sich mit dem kognitivistischen, akzentuiert aber die Rolle des Lerners und seines Wissenshintergrundes beim Verarbeiten von Feedbackmitteilungen. Nach Mory (2004) bietet er einen Erklärungsrahmen für die nicht uniforme Verarbeitung von Feedback bzw. für die Frage, wie Feedback wirkt und wie oder warum es nicht wirkt. Insgesamt ist der konstruktivistische Ansatz für die Feedbackforschung, vor allem für die konkrete praktische Umsetzung, aber noch wenig ausgearbeitet. Insbesondere für eine computergestützte Lernumgebung stellt sich die Herausforderung, dass diese im konstruktivistischen Sinn entsprechend viele Freiheitsgrade zu berücksichtigen hat

(Musch, 1999), um den individuellen Voraussetzungen der Lerner und den verschiedenen Anforderungen der jeweiligen Feedbackinterventionen gerecht zu werden.

Basierend auf den vorangegangenen Ausführungen zum Feedbackkonzept in Bezug auf die Lerntheorien liegt dieser Arbeit ein kognitivistisches Verständnis von Feedback zugrunde. Feedback wird also als Information verstanden, die einem Lerner bezüglich bestimmter Aspekte seiner Leistung gegeben wird. Der Korrektur von Fehlern wird dabei im Vergleich zur Bestätigung von korrekten Antworten ein größeres Gewicht beigemessen. Die Bedeutung des Wissenshintergrunds des Lerners und der möglicherweise stark individuellen konstruktiven Leistungen beim Verarbeiten von Feedback ist im Rahmen der Diskussion von Feedbackeffekten zu berücksichtigen.

3.2 Merkmale der Feedbackgestaltung

Auch wenn der Begriff Feedback seinem Prinzip nach immer dasselbe meint, ist es dennoch kein einheitliches Phänomen (Bangert-Drowns et al., 1991). Es handelt sich vielmehr um einen Oberbegriff für eine Vielzahl von Rückmeldungsarten, die auf mehreren Dimensionen beschrieben und voneinander unterschieden werden können (Shute, 2008). Hierzu werden in weitestgehender Überstimmung primär der Inhalt, der Zeitpunkt und der Präsentationsmodus einer Rückmeldung gezählt (Kulhavy & Stock, 1989; Mory, 2004; Shute, 2008). Daneben werden Feedbacks gelegentlich auch hinsichtlich der durch sie anvisierten Prozesse unterschieden (Bangert-Drowns et al., 1991; Narciss & Huth, 2004). Die Dimensionen können variabel kombiniert werden, woraus die breite Vielfalt von Feedbackarten, die in der Literatur zu finden ist, resultiert.

Rückmeldungen werden immer in einem bestimmten Kontext gegeben, den es für die Auswahl und die konkrete Ausgestaltung einer Feedbackart zu berücksichtigen gilt. Narciss und Huth (2004) formulieren, dass der „informative Wert von Feedback“ (S. 183, *Übersetzg. v. Verf.*) von situativen Faktoren (z.B. Lernziel, inhaltliche Anforderungen der Aufgaben, mögliche Quellen für typische Fehler) sowie bestimmten Merkmalen des Lerners (z.B. Vorwissen, vorhandene Fähigkeiten, Lernziel) mitbestimmt wird (siehe Abbildung 1). Diese Auffassung kontextueller Einflüsse auf die Gestaltung und Wirkung

von Rückmeldungen bildet den Hintergrund, vor dem die folgenden Beschreibungen der zentralen Dimensionen bzw. Merkmale der Feedbackgestaltung zu verstehen sind.

3.2.1 Feedbackinhalt

Mit dem Begriff Feedbackinhalt ist die Art der Information gemeint, die in einer Rückmeldung enthalten ist. Dem Inhalt wird hinsichtlich der Wirksamkeit von Feedback für das Lernen außerordentlich viel Bedeutung in der Literatur beigemessen. Dies ist wiederum vor dem Hintergrund des kognitivistischen Paradigmas in Abgrenzung zum operanten Ansatz zu sehen: Wenn Feedback die Information ist, die vom Lerner zum Korrigieren fehlerhafter oder unzureichender Antworten genutzt wird, sollte sich die Art und Menge der rückgemeldeten Informationen entscheidend auf den Lern-/ Leistungsprozess auswirken (Kulhavy & Stock, 1989).

Die Art und Menge rückgemeldeter Informationen werden im Rahmen der Feedbackforschung auch unter dem Begriff der Komplexität behandelt (Dempsey, Driscoll & Swindell, 1993; Shute, 2008). Danach lassen sich weniger umfangreiche Arten der Rückmeldung von komplexeren Arten unterscheiden. Kulhavy (1977) führt hierzu aus:

> If we are willing to treat feedback as a unitary variable, we can then speak of its form or composition as ranging along a continuum from the simplest 'Yes-No' format to the presentation of substantial corrective or remedial information that may extend the response content, or even add new material to it. Hence, as one advances along the continuum, feedback complexity increases until the process itself takes on the form of new instruction, rather than informing the student solely about correctness. (S. 212)

Neben der Komplexität von Rückmeldungen wird häufig auch deren Spezifität betrachtet. Mit Feedbackspezifität ist das Informationslevel einer Mitteilung gemeint (Goodman, Hendrickx & Wood, 2004), das heißt das Ausmaß, in dem die Rückmeldung auf konkrete Antworten, relevante Verhaltensweisen oder Prozesse bezogen ist (Shute, 2008).

Komplexität und Spezifität stellen also zwei Dimensionen dar, anhand derer sich Feedbackmitteilungen in einem ersten Schritt charakterisieren lassen. Feedbackinhalte können mehr oder weniger komplex und mehr oder weniger spezifisch sein. Ein strenger linearer Zusammenhang beider Merkmale kann allerdings nicht unterstellt werden. Die Spezifität einer Intervention muss nicht mit der Menge an der beinhalteten Information ansteigen. Sehr komplexes Feedback muss nicht zwangsläufig hoch spezifisch sein, et

vice versa. Die Beschreibung von Feedbackinhalten anhand ihrer Komplexität und Spezifität ist also eher als eine grobe Annäherung anzusehen.

Zur Kategorisierung der vielfältigen inhaltlich unterscheidbaren Feedbackarten sind verschiedene Vorschläge vorgelegt worden (z.B. Clariana, 2001 zitiert nach Narciss, 2006; Dempsey, Driscoll & Swindell, 1993; Kulhavy & Stock, 1989; Mason & Bruning, 2001; Narciss, 2006; Schimmel, 1988; Shute, 2008). Weitestgehender Konsens besteht in der Klassifizierung der Feedbackarten mit begrenztem Informationsgehalt, zu denen die folgenden gezählt werden: *Knowledge of Result*, *Knowledge of Correct Result* und *Try-again Feedback* (vgl. Tabelle 1 für Zusammenfassung).

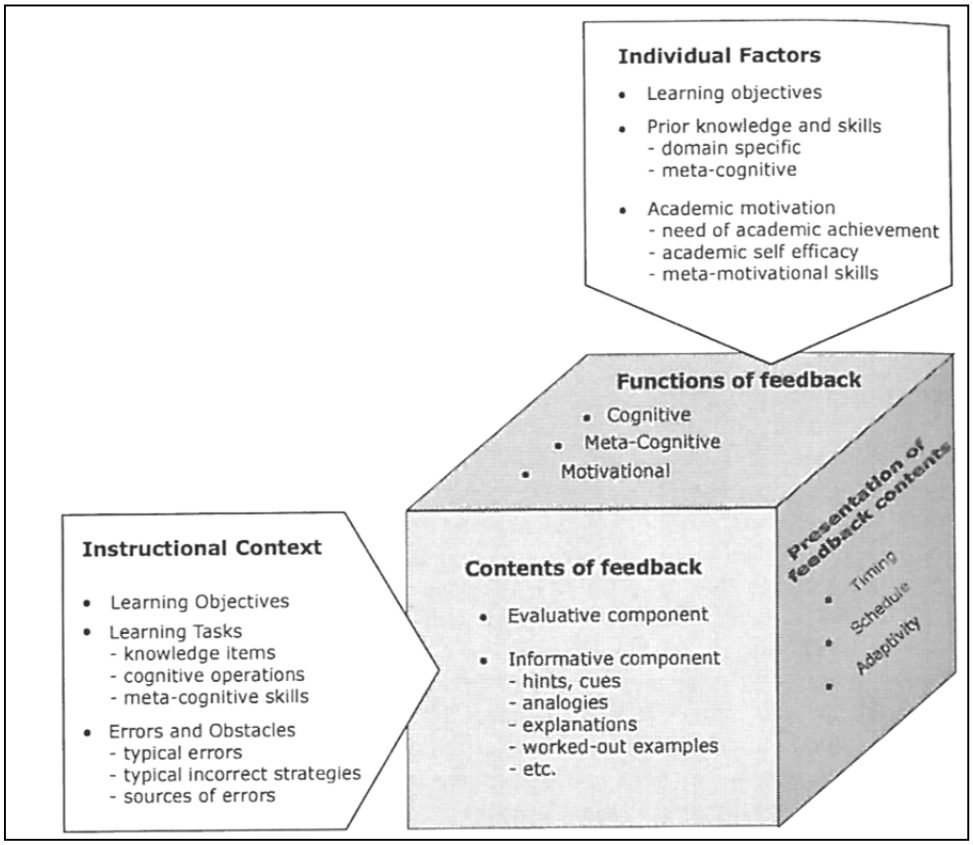

Abbildung 1 Determinanten des Informationswerts von Feedback (Narciss & Huth, 2004, S. 184).

Das Feedback Knowledge of Result teilt mit, ob eine Antwort bzw. eine Leistung richtig oder falsch ist. Hierbei handelt es sich um die grundlegendste und gleichzeitig inhaltlich am wenigsten umfangreiche Art der Rückmeldung. Auch in der deutschsprachigen

Literatur wird typischerweise auf die englische Bezeichnung zurückgegriffen. Es finden sich aber auch alternative Begriffe zu Knowledge of Result. Häufiger ist noch der Begriff Verifikation zu finden, vereinzelt auch die Bezeichnungen *Accuracy Feedback* (z.B. in Beckmann, N., Beckmann & Elliott, 2009) oder *Informational Feedback* (z.B. in Rakoczy, Klieme, Bürgermeister & Harks, 2008).

Unwesentlich komplexer als Knowledge of Result ist das Feedback Knowledge of Correct Result, das nicht explizit mitteilt, ob eine abgegebene Antwort richtig oder falsch ist, sondern es nennt lediglich die richtige Lösung für eine Aufgabenstellung. Durch den Vergleich der rückgemeldeten Lösung mit der eigenen Antwort wird dem Lerner ersichtlich, ob er richtig oder falsch geantwortet hat.

Tabelle 1 Feedbackarten mit begrenztem Informationsgehalt (Narciss, 2006, S. 19 u. S. 23; Shute, 2008, S. 160)

Feedbackart	Beschreibung	Gestaltungsbeispiele
Knowledge of Result (synonym: Verifikation, Accuracy Feedback)	Informiert, ob eine Antwort richtig oder falsch ist; Kann auch summativ, d.h. für eine Menge von Antworten, gegeben werden	➢ „Das ist richtig.“/„Das ist falsch.“ ➢ „Stimmt“/„Stimmt nicht“ ➢ „2 von 5 Aufgaben sind richtig.“/„25% der Antworten sind nicht richtig.“
Knowledge of Correct Result	Nennt die richtige Lösung für eine Aufgabenstellung	➢ „Die richtige Antwort lautet …“. ➢ „Korrekt ist Antwortalternative ‚b‘.“
Try-again Feedback bzw. Repeat-until-Correct Feedback	Informiert, dass eine Antwort falsch ist, und gewährt weiteren Antwortversuch bzw. so viele, bis die richtige Lösung erreicht wird	➢ „Das ist falsch. Versuch es noch einmal.“ ➢ „Das ist falsch.“, die Aufgabe wird dann erneut vorgegeben

Daneben wird noch das Try-again Feedback bzw. *Repeat-until-Correct Feedback* unterschieden. Dabei wird dem Lerner nach einer falschen Antwort mitgeteilt, dass die Lösung nicht korrekt ist, woraufhin er zu einem erneuten Antwortversuch aufgefordert wird. In manchen Untersuchungen ist die Anzahl erneuter Antwortversuche nicht beschränkt, es können tatsächlich so viele Antworten gegeben werden, bis die richtige

Lösung erreicht ist. In anderen Fällen wird hingegen nur eine maximale Anzahl an Antwortversuchen pro Aufgabe zugelassen. Gestaltungsmöglichkeiten für die genannten drei Arten des Feedbacks mit begrenztem Informationsgehalt sind in Tabelle 1 aufgeführt.

Die genannten Feedbacks Knowledge of Result, Knowledge of Correct Result und Try-again Feedback werden häufig auch als „einfache“ Feedbackarten zusammengefasst und der Gruppe der so genannten elaborierten Arten gegenübergestellt. Der Begriff Elaboration kann im Rahmen der Feedbackforschung auf Kulhavy und Stock (1989) zurückgeführt werden. Sie schlugen vor, dass eine Rückmeldung, die über Knowledge of Result hinausgeht, also in irgendeiner Weise ergänzende Information beinhaltet, ihrem Inhalt nach elaboriert ist. Nach diesem Verständnis werteten die Autoren auch Knowledge of Correct Result und Try-again Feedback als elaborierte Feedbackarten. Doch in aller Regel werden diese inzwischen als einfache, nicht als elaborierte Feedbackarten, eingeordnet (Narciss, 2006; Shute, 2008). Das Verständnis des Begriffs Elaboration selbst hat sich dagegen nicht geändert. Als elaboriertes Feedback werden Inhalte verstanden, die in irgendeiner Weise ergänzende, erklärende Informationen bieten, wobei sie aber durchaus auch Knowledge of Result oder Knowledge of Correct Result integrieren können.

Dabei bleibt eine gewisse Ungenauigkeit des Begriffs Elaboration, der auch Spielraum für die konkrete Ausgestaltung von Rückmeldungen bietet und in der Tat ist die Vielzahl der in der Literatur auffindbaren Feedbackinterventionen aus dem Bereich der elaborierten Rückmeldungen. Die Klassifikationssysteme (z.B. Clariana, 2001 zitiert nach Narciss, 2006; Dempsey, Driscoll & Swindell, 1993; Kulhavy & Stock, 1989; Mason & Bruning, 2001; Narciss, 2006; Schimmel, 1988; Shute, 2008) stimmen hier, bei der Kategorisierung elaborierter Rückmeldungen, auch nicht mehr überein. Zum Teil werden verschiedene Aspekte in den Vordergrund gestellt (Narciss, 2006): funktionale (Chi, 1996; Clariana 2001, zitiert nach Narciss, 2006), inhaltliche (Kulhavy & Stock, 1989; Mason & Bruning, 2001; Narciss, 2006; Shute, 2008) oder eine Kombination beider (Schimmel, 1988). Zum Teil sind die Kategoriensysteme auch nur unterschiedlich differenziert. Die neueren Einteilungen von Shute (2008) und Narciss (2006) integrieren überdies explizit Teile vorausgegangener Klassifikationen und bieten damit die umfangreichsten und differenziertesten Unterteilungen inhaltlich elaborierter Feedbackarten. Trotzdem wird keinem der bestehenden Klassifikationsansätze in der Literatur der Vorrang gegeben; eine einheitlich verwendete Systematisierung liegt nicht

vor. In Tabelle 2 sind verschiedene, an inhaltlichen Aspekten orientierte Klassifikationssysteme für elaborierte Feedbackarten zusammengefasst, wobei versucht wurde, gleiche oder ähnliche Kategorien verschiedener Klassifikationen auf derselben Höhe anzuordnen.

Kurz zusammengefasst lässt sich sagen, dass sich elaboriertes Feedback auf die behandelte Thematik, die Aufgabenstellung oder die konkreten Antwort beziehen kann (Mason & Bruning, 2001; Shute, 2008). Dabei können bestimmte Aspekte hervorgehoben, Informationen ergänzt und/oder erklärt werden. Ferner kann sich elaboriertes Feedback an der Fähigkeit oder dem Prozess, der zu einer Antwort geführt hat bzw. an der gezeigten Leistung selbst orientieren und sich dabei auf den Fehler oder entgegengesetzt auf die erwünschte Antwort („die Lösung“) konzentrieren. Besonders hier wird deutlich, dass ein und dieselbe Feedbackart in der konkreten Anwendung unterschiedlich detailliert ausgestaltet sein kann. Beispielsweise kann ein Fehler einerseits lediglich benannt oder andererseits auch erklärt werden. Statt einzelnen Fehlern können auch systematisch falsche Konzepte von Prozessen oder Sachverhalten, so genannte *Misconceptions*, behandelt werden. Soll der Lösungsprozess unterstützt werden, können Strategien, die ganz spezifisch auf die jeweilige Aufgabe ausgerichtet sind oder Strategien genereller Natur eingesetzt werden. Einige Möglichkeiten der Ausgestaltung der genannten elaborierten Rückmeldungen finden sich bei Shute (2008) und Narciss (2006) und sind Tabelle 3 integriert.

Die konkrete Ausgestaltung der elaborierten Feedbackarten hängt natürlich auch stark vom Kontext der Lern- bzw. Leistungssituation, insbesondere der vorliegenden Domäne, ab. Ein anschauliches Beispiel ist hier fehlerbezogenes Feedback: Während in der Mathematik Fehler eineindeutig identifiziert und analysiert werden können, sind falsche oder unzureichende Antworten im Rahmen der Überprüfung des Textverständnisses weniger leicht und eindeutig zu bestimmen. Prinzipiell sind alle der genannten elaborierten Feedbackarten für den Bereich des Textverstehens/der Lesekompetenz anwendbar, wobei eine gewisse Präferenz zur Vermittlung (meta-)kognitiver Hinweise (das „Know how“) zur Bewältigung einer Anforderung erkennbar ist (siehe Abschnitt 3.5.1.1).

Tabelle 2 Klassifikationen elaborierter Feedbackarten

Schimmel, 1988 (S. 184-186)	**Kulhavy & Stock, 1989** (S. 285-287)	**Mason & Bruning, 2001** (Abs. 8)	**Narciss, 2006** (S. 23)	**Shute, 2008** (S. 160)
	Task-specific (Gives mostly the correct answer)			
			Knowledge on Task Constraints (Provides information about task characteristics/constraints, sub-tasks or processing rules)	
with General Review (Contains summary statements of the instructional content that preceded learner's wrong answer)	**Instruction-based** (Contains information derived from the specific lesson material being studied, but not directly from the actual task completed prior to feedback)	**Topic-contingent** (Addresses the target topic by returning learner to passages/other learning material where the correct information is located or by giving additional information)		**Topic Contingent** (Provides information relating to the target topic currently being studied)
Attribute-isolation (Reviews the attributes/ characteristics of a concept on which the student made an error)		**Attribute-isolation** (Provides verification and highlights central attributes of the target concept)	**Knowledge about Concepts** (Provides conceptual knowledge relevant to the task)	**Attribute Isolation** (Addresses central attributes of the target concept/skill being studied)
	Extra-instructional (Adds information from outside the immediate lesson environment)	**Response-contingent** (Focus on specific response; may describe why correct answer is correct/why the incorrect answer is wrong)		**Response Contingent** (Focus on specific response; may describe why correct answer is correct/why the incorrect answer is wrong, no formal error analysis)
with Specific Review (A step-by-step solution to an incorrectly answered problem, without the final step)			**Knowledge on How to Proceed** (Provides strategic knowledge that is relevant to task solution) **Knowledge on Meta-Cognition** (Provides information that can stimulate regulatory learning processes)	**Hints/Cues/Prompts** (Guides in the right direction, e.g., strategic hint on what to do next; does not present correct answer)
Bug-related (Aims at correcting a faulty mental model of a procedure) (Bugs are systematic errors)		**Bug-related** (Provides verification and addresses specific errors, can assist in identifying procedural errors)	**Knowledge about Mistakes** (Addresses specific error)	**Bugs/Misconceptions** (Provides information about specific errors or misconceptions)
				Informative Tutoring (Combines verification, error flagging, and strategic hints on how to proceed)

Tabelle 3 Mögliche Inhalte elaborierter Feedbackarten (nach Narciss, 2006, S. 23, erweitert um Shute, 2008, S. 160)

Elaborierte Feedbackart	Gestaltungsbeispiele
Knowledge on Task Contraints	➢ Hinweise auf Art der Aufgabe bzw. Aufgabenanforderungen, ➢ Hinweise auf Bearbeitungsregeln, ➢ Hinweise auf Teilaufgaben
Topic Contingent Feedback	➢ Verweis auf relevante Passagen, ➢ *Reteaching*, ➢ Geben zusätzlicher/neuer Informationen zum Thema
Knowledge about Concepts	➢ Hinweise auf Fachbegriffe, ➢ Beispiele für Begriffe, ➢ Hinweise auf Begriffskontext, ➢ Erklärungen zu Begriffen
Response Contingent Feedback	➢ Erklärung, warum die gewählte Antwort richtig ist bzw. warum die richtige Antwort korrekt ist, ➢ Erklärung, warum die gewählte Antwort falsch ist bzw. warum die falsche(n) Antwortalternative(n) falsch ist (sind)
Hints, Cues, Prompts	➢ Fehlerspezifische Korrekturhinweise, ➢ Aufgabenspezifische Lösungshinweise, ➢ Hinweise auf (meta-)kognitive Lösungsstrategien, ➢ (metakognitive) Leitfragen, ➢ Lösungsbeispiele (w*orked-out examples*)
Knowledge about Mistakes/Bugs or Misconceptions	➢ Anzahl der Fehler, ➢ Ort des Fehlers/der Fehler, ➢ Art oder Ursache des Fehlers/der Fehler Anmerkung: Hier gelten nicht nur Fehler, die in einer einzelnen Aufgabe gemacht werden, sondern auch systematische Fehler (*bugs*) bzw. falsche Konzepte.

Alle erläuterten Feedbackarten sind eigenständig einsetzbar. Daneben können sie aber auch in Kombination, zeitgleich (z.B. in Kulhavy, White, Topp, Chan & Adams, 1985) oder im Sinne graduierter Hilfen schrittweise (z.B. in Narciss & Huth, 2004), dargeboten werden. Ein Beispiel für eine schrittweise Feedbackgabe ist das „informative tutorielle Feedback“ von Narciss und Huth (2004) für den Bereich der Mathematik. Mit jedem falschen Antwortversuch für eine Aufgabe erhält der Lerner zunehmend umfangreicheres Feedback. Die maximalste Rückmeldung für eine Aufgabe präsentiert Knowledge of

Result, benennt den Fehler in der Antwort und bietet zusätzlich strategische Hinweise zur Lösung der Aufgabe an, alles in ein und derselben Mitteilung.

Die erfolgte Darstellung und Beschreibung möglicher Feedbackinhalte unterstellt implizit, dass die Rückmeldungen auf der Ebene kognitiver und metakognitiver Prozesse beschrieben sind. Das heißt im Normalfall spiegeln alle Feedbackarten, die einfachen und die elaborierten, in ihrer Formulierung (meta-)kognitive Prozesse wider. Davon abzugrenzen sind Mitteilungen, die allein auf die Motivation des Lerners ausgerichtet sind, indem beispielsweise mehr Anstrengung zur Bewältigung einer Aufgabe verlangt oder eine richtige Lösung gelobt bzw. auf entsprechende Bemühungen zurückgeführt wird. Davon unberührt ist natürlich der Umstand, dass jede Feedbackmitteilung sich sowohl auf der kognitiven als auch der motivationalen Ebene des Lerners auswirken kann (siehe Abschnitt 3.3.).

Bezugsnormorientierung: Die Ausgestaltung von Rückmeldungen wird auch durch die Wahl der Bezugsnorm bestimmt, die gewissermaßen quer zu der oben erläuterten inhaltlichen Einteilung von Feedbackarten liegt. Es können eine kriterienbezogene, eine intraindividuelle und eine interindividuelle Bezugsnorm unterschieden werden (Krause, 2007).

In der großen Mehrheit der Feedbackstudien wird die kriterienbezogene Bezugsnorm angewendet, indem der Feedbackinhalt am Lern- oder Leistungsziel orientiert ist. Das heißt, für jede Aufgabenstellung kann festgestellt werden, was als richtige Lösung gilt und was im Gegensatz dazu eine falsche oder unzureichende Antwort ist. Feedback mit einer intraindividuellen Bezugsnorm bezieht sich auf die individuelle Entwicklung eines Lerners in einem Leistungsbereich (z.B. in Korsgaard & Diddams, 1996; Schunk & Rice, 1991) oder ist auch „nur“ auf der persönlichen Ebene formuliert (z.B. in Schunk & Rice, 1986). Hierzu sind demnach auch Feedbacks zu zählen, die den Lerner explizit auf der Ebene der Motivation ansprechen (z.B. „Streng dich an.“ oder „Das hast du schön gemacht.“). Im Rahmen der interindividuellen Bezugsnorm wird einem Lerner rückgemeldet, wie gut seine gezeigte Leistung oder Fähigkeit im Vergleich zu anderen Lernern ist (z.B. in Cervone & Wood, 1995).

3.2.2 Zeitpunkt der Feedbackpräsentation

Jedes Feedback ist neben inhaltlichen Gesichtspunkten auch hinsichtlich des Zeitpunktes seiner Darbietung beschreibbar. Dabei kann es entweder unmittelbar, also nach der Beantwortung einer Aufgabe, oder verzögert gegeben werden. „Verzögert“ meint nach der Taxonomie von Dempsey und Wager (1988), dass die Rückmeldung erst nach einem festgelegten Zeitintervall nach Abschluss einer Aufgabenstellung präsentiert wird (z.B. in Dihoff, Brosvic & Epstein, 2003). Häufig wurde in Untersuchungen eine Verzögerung von entweder ein paar Sekunden oder ein bis zwei Tagen gewählt (Kulik & Kulik, 1988). Die Festlegung, wann es sich um verzögertes Feedback handelt, wird allerdings oft in Relation zur Vergleichsbedingung der unmittelbaren Feedbackgabe bestimmt (Shute, 2008). So kann es sich in einer Untersuchung um ein verzögertes Feedback handeln, wenn es am Ende des Tests statt unmittelbar nach jeder Testantwort gegeben wird (z.B. in Morrison, Ross, Gopalakrishnan & Casey, 1995). In einer anderen Untersuchung kann die Feedbackgabe am Ende des Tests aber auch eine nicht-verzögerte Bedingung darstellen, wenn sie mit einer Gruppe verglichen wird, die Feedback einen Tag nach Abschluss des Tests erhält (z.B. Kulhavy & Anderson, 1972).

Der Zeitpunkt der Feedbackpräsentation ist Gegenstand einer Reihe älterer Untersuchungen, die sich auf den so genannten *Delay-Retention Effect* beziehen. Dieser Begriff steht für das Phänomen, dass die Lernleistung durch verzögertes im Vergleich zu unmittelbarem Feedback erhöht sein kann (Kulhavy & Anderson, 1972). Die besondere Relevanz dieser Untersuchungen erschließt sich, wenn man bedenkt, dass die frühen Feedbackstudien unter dem operanten Paradigma durchgeführt wurden, für das die Unmittelbarkeit der Rückmeldung (d.h. des Verstärkers) ein grundlegendes Prinzip darstellt (Mory, 2004). Dementsprechend groß war die Aufmerksamkeit, die dem Phänomen anfangs entgegengebracht wurde. Doch letztlich stellt sich die Befundlage so dar, dass die Vorteile verzögerten Feedbacks auf wenige, spezielle Bedingungen beschränkt sind (Kulik & Kulik, 1988; vgl. Abschnitt 3.5.1.2).

3.2.3 Präsentationsmodus von Feedback

Feedback kann sowohl durch Personen (z.B. Lehrer, Testleiter, Peer) als auch „unpersönlich“ dargeboten werden (Hattie & Timperley, 2007). Letzteres geschieht, wenn Rückmeldungen in einer Lern- bzw. Testumgebung integriert sind und ohne

unmittelbares Zutun einer außenstehenden Person darüber präsentiert werden. Dieses Vorgehen ist grundsätzlich auch in papierbasierten Lern- oder Testmaterialien umsetzbar (z.B. in Dihoff, Brosvic, Epstein & Cook, 2004; Kulhavy et al., 1985). Aufgrund der technischen Entwicklungen wird es inzwischen aber vorrangig in computerbasierten Programmen angewendet, die zudem häufig das Szenario der Wahl in den Feedbackstudien der letzten Jahre darstellen.

Wird Feedback durch eine Person gegeben, geschieht dies typischerweise in mündlicher Form. Dabei wird die reine Feedbackinformation automatisch durch prosodische Merkmale (vgl. Bußmann, 2002, S. 542) wie Betonung, Tonfall oder Pause eingefärbt. Darüber hinaus kommen bei der persönlichen Interaktion normalerweise auch nonverbale Signale des Feedbackgebers, seine Mimik und Gestik, zum Tragen. Dadurch erhält der Lerner zusätzliche Informationen, die ihm das Verständnis der Rückmeldung möglicherweise erleichtern. Wie stark diese Aspekte in der konkreten Untersuchungssituation jedoch zugelassen werden, hängt von den Zielen und der Anlage der jeweiligen Untersuchung ab. Deutlich zum Tragen kommen sie typischerweise in Interventionen, in denen ein Lerner beim Bearbeiten einer Aufgabenstellung durch einen kompetenten Anderen (auch „Tutor“ genannt) angeleitet und unterstützt wird. Der Dialog zwischen den Beteiligten formt dabei die Grundlage der Intervention und Feedback, aber auch andere Hilfestellungen stellen einen wichtigen Baustein dar.

Tutoring-Szenarien sind nicht selten im Rahmen der Feedbackforschung vorzufinden (z.B. Alber-Morgan, Matheson Ramp, Anderson & Martin, 2007; Azevedo, 2007; Azevedo, Cromley & Seibert, 2004; Pany, McCoy & Peters, 1981; Schunk & Rice, 1991; Winne, Graham & Prock, 1993). Durch die mehr oder weniger ausgeprägte Interaktion zwischen Lerner und Tutor bzw. Testleiter, einschließlich des mündlichen Feedbacks, wird die gesamte Untersuchungssituation allerdings in bestimmter Weise geprägt. Einerseits könnte dieses soziale Setting auf Seiten des Lerners eine höhere Verbindlichkeit (*Commitment*) oder Ernsthaftigkeit bewirken als beispielsweise ein stark künstliches, nüchternes Laborexperiment. Andererseits ist die Intervention aber auch weniger standardisierbar, wodurch die Generalisierbarkeit der Ergebnisse eingeschränkt wird.

Akustisches Feedback ist aber nicht nur in Verbindung mit einer Person als Feedbackgeber einsetzbar, sondern es kann prinzipiell auch über computergestützte Lern- bzw. Testumgebungen übermittelt werden, zum Beispiel mithilfe eines implementierten

Sprachgenerators (z.B. in Farmer, Klein & Bryson, 1992) oder mittels Audiofiles. Im Normalfall wird beim Einsatz von computer- und auch papiergestützten Lern- bzw. Testumgebungen aber auf schriftliches Feedback zurückgegriffen. Es ist die wohl häufigste Präsentationsform in Feedbackstudien im pädagogisch-psychologischen Bereich.

Typischerweise impliziert die Gabe von Feedback über eine Lern- oder Testumgebung, dass der Lerner selbstständig, also ohne direkte Interaktion mit einer anderen Person, arbeitet. Das bedeutet unter anderem auch, dass die Möglichkeit nach direkten Rückfragen, etwa zum Verständnis des Feedbacks, meist nicht mehr besteht. Im Vergleich zur personengebundenen Feedbackgabe ist zu bedenken, dass das selbstständige Arbeiten mit Feedback in einer automatisierten, computergestützten Umgebung möglicherweise das Commitment seitens des Lerners reduziert und höhere Anforderungen an seine (Test-)Motivation stellt. Dafür ist die „unpersönliche" Feedbackgabe aber auch maximal standardisierbar. Die Fehleranfälligkeit ist vor allem bei komplexen, gestuften Feedbacksystemen minimal und der Zeitpunkt der Feedbackgabe kann, sofern von Interesse, in computergestützten Programmen exakt eingehalten werden (z.B. für verzögertes Feedback mit festgelegtem Intervall zwischen Antwort des Lerners und Rückmeldung). Darüber hinaus handelt es sich im Normalfall auch um die ökonomischste Variante einer Untersuchungsanlage.

Im Gegensatz zu den schriftlichen oder akustischen Darstellungsformen sind die Möglichkeiten, Feedback in Form von Bildern einzusetzen, deutlich eingeschränkt. Die Rückmeldung, dass eine Antwort richtig ist, kann anstatt mit Worten („richtig" o.ä.) beispielsweise auch mit einem lachenden Gesicht (z.B. Ball, Hoyle & Towse, 2010), einem Häkchen oder einem „Daumen hoch"-Symbol mitgeteilt werden. Dass ein Fehler gemacht wurde, kann entsprechend über ein trauriges Gesicht (z.B. Ball et al., 2010) oder ähnliches dargestellt werden. Das Prinzip des Knowledge of Correct Result ist bei Dekodieraufgaben oder einfachen Merkaufgaben mit dem Zeigen von entsprechenden Bildern der gesuchten Information umsetzbar (z.B. Solman & Wu, 1995). Von besonderem Vorteil könnte diese Darstellung beispielsweise für Leseanfänger sein, weil damit der Leseaufwand geringer gehalten wird, oder die Bilder sind gestalterisches Mittel für eine aufgelockerte Gestaltung der Feedbackgabe in einer Lern- bzw. Testumgebung. Aber auch elaborierte Feedbacks sind im Prinzip bildlich/grafisch darstellbar. Beispiele dafür sind in spezifischen Kontexten wie der Konstruktion räumlicher Situationsmodelle

zu finden, für die eine Landkarte oder ein Raumplan als Hilfestellungen eingesetzt werden (z.B. Langer, Keenan & Schreiner, 1995). In ähnlicher Weise werden zur Unterstützung des Verständnisses von Textstrukturen *Maps* als Rückmeldungen verwendet, die die geforderten Textinformationen und ihre Relationen zueinander grafisch darstellen (z.B. Yang, Yeh & Wong, 2008). Ebenso sind Anwendungen bildlichen/grafischen Feedbacks bei Anforderungen der Text-Bild-Integration möglich (vgl. Bodemer, Ploetzner, Feuerlein & Spada, 2004; Brünken, Seufert & Zander, 2005), indem das Verständnis eines Sachverhalts auch durch Hervorhebungen im Bild unterstützt werden kann. Insgesamt wird die bildliche Darstellungsform von Feedback aber selten verwendet.

Die Art und Weise, wie Feedback präsentiert wird, lässt sich also auf zwei Ebenen beschreiben: die Form der Darstellung (schriftlich, bildlich/grafisch oder akustisch) und der „Übermittler“ (eine Person oder die Lern-/Testumgebung). Beide Ebenen sind stets konfundiert. Insgesamt stellt sich die Forschungsliteratur so dar, dass der Präsentationsmodus von Feedback selten die interessierende Untersuchungsvariable ist. Vielmehr resultiert die Form der Feedbackpräsentation aus der Testumgebung, die man schaffen möchte. Die Besonderheiten einer Darstellungsform und mögliche Konfundierungen sind dann aber bei der Interpretation und Diskussion der Befunde zu berücksichtigen.

Zuletzt kann noch eine Art Sonderfall der Feedbackvermittlung erwähnt werden, und zwar eine implizite Vermittlung von Knowledge of Result. Dabei wird die Information, dass eine Antwort richtig oder falsch ist, nicht explizit schriftlich, bildlich oder mündlich präsentiert, sondern implizit über den Ablauf einer (normalerweise computerbasierten) Lern- bzw. Testumgebung (Shute, 2008). Wird dem Lerner nach seiner Antwort auf eine Aufgabenstellung beispielsweise die nächste Aufgabe oder „Seite“ im computerbasierten Programm gezeigt, impliziert dies, dass die abgegebene Antwort richtig gewesen ist. Im Gegensatz dazu kann bei einer falschen Antwort dieselbe Aufgabenstellung wiederholt präsentiert werden, ohne irgendeine explizite Mitteilung. Die Gestaltungsform der impliziten Rückmeldung setzt natürlich voraus, dass die Lerner im Vorfeld entsprechend instruiert werden. Es ist sicherlich auch nicht unstrittig, ob es sich hier überhaupt um Feedback im eigentlichen Sinn handelt bzw. wie (gut) sich diese Form in den Kanon der Feedbackarten integrieren lässt. Nichtsdestotrotz finden sich Anwendungsbeispiele für

implizites Feedback, so zum Beispiel auch im Rahmen von Maßnahmen zur Förderung des lauten, flüssigen Lesens (z.B. Alber-Morgan et al., 2007; Pany et al., 1981). Während Probanden einen Text laut vorlesen, werden sie bei falsch ausgesprochenen Wörtern vom anwesenden Feedbackgeber unterbrochen und explizit korrigiert (z.B. mit Knowledge of Correct Result oder Knowledge of Result), werden sie beim Lesen nicht unterbrochen, bedeutet das für sie, dass sie die Wörter richtig ausgesprochen haben.

3.3 Funktionen von Feedback

Eine Funktion des Feedbacks besteht darin, nach richtigen Antworten oder gewünschtem Verhalten diesen (Leistungs-)Stand zu spiegeln (Mory, 2004). Die wesentliche Funktion von Feedback wird jedoch in der Korrektur von Fehlern gesehen (Kulhavy, 1977) und dafür kann es in der Art eingesetzt werden, dass es die Diskrepanz zwischen einem aktuellen Leistungsstand und einem gesetzten Zielzustand signalisiert (Hattie & Timperley, 2007).

Diese Signalfunktion hat das einfachste Feedback Knowledge of Result, das allein oder als Komponente eines elaborierten Feedbacks gegeben werden kann. Es wird angenommen (vgl. Shute, 2008), dass das Rückmelden der Diskrepanz die Unsicherheit bezüglich der Korrektheit der Antwort reduziert, was wiederum zu vermehrter Anstrengung und der Anwendung effizienterer Strategien zur Bewältigung einer Aufgabe führen kann (Song & Keller, 2001). Solch ein selbstregulierter Lösungsprozess erscheint sehr voraussetzungsvoll, sofern die Aufgabenstellung nicht ausschließlich einer Merkleistung bedarf, sondern eher Verstehen voraussetzt, und der Fehler auch tatsächlich auf falschem oder unzureichendem Verstehen beruht und nicht etwa ein Flüchtigkeitsfehler ist.

Als weniger voraussetzungsvoll in der Umsetzung kann das elaborierte Feedback angesehen werden, das per Definition weiterführende, das Verständnis bzw. den Wissenserwerb unterstützende Informationen enthält und so der Korrektur von Fehler nützlich sein kann. Nach Butler und Winne (1995) kann Feedback im Informationsverarbeitungsprozess zum einen in der Art wirken, dass es Gedächtnisinhalte ergänzt, überschreibt, präzisiert oder restrukturiert. Zum anderen kann es (meta-) kognitive Prozesse stimulieren. Die vermittelten Informationen können dabei domänenspezifisch oder übergreifender Natur sein.

Shute (2008) zählt neben der Signalfunktion und der korrigierenden Funktion von Feedback noch eine dritte mögliche Wirkung von Feedback auf: Es kann die Arbeitsgedächtnisbelastung des Lerners reduzieren. Eine hohe Arbeitsgedächtnisbelastung, die insbesondere im Rahmen komplexer Anforderungen auftritt, provoziert vor allem bei geringen Vorkenntnissen oder bei Schwierigkeiten in einer Domäne schlechtere Leistungen (Paas, Renkl & Sweller, 2003, 2004). Um die Arbeitsgedächtnisbelastung beim Lerner zu senken, wird auch im Rahmen von Feedbackinterventionen auf ausgearbeitete Beispiele (worked-out examples; Sweller, van Merriënboer & Paas, 1998) oder bei geringen Vorkenntnissen in einem Wissensbereich auch auf elaborierten Erklärungen/Erläuterungen (Moreno, 2004) zurückgegriffen.

Feedbacks werden also hauptsächlich kognitive und metakognitive Funktionen zugeschrieben. Davon schwer zu trennen ist, dass eine Feedbackintervention an sich sehr wahrscheinlich ebenso Auswirkungen auf die Motivation des Lerners hat (vgl. Kluger & DeNisi, 1998). Desgleichen setzt die Wirkung eines Feedbacks immer auch die entsprechende Motivation bzw. Bereitschaft des Lerners voraus, die Informationen zu nutzen und umzusetzen (Kulhavy, 1977).

3.4 Indikatoren der Wirkung von Feedback

Jede Feedbackstudie basiert auf einer Treatmentphase, in der im Zuge der Bearbeitung von Aufgaben Feedback gegeben wird. Um die Effekte der Feedbackintervention auf die Leistung zu überprüfen, wird für gewöhnlich auf ein Posttestdesign zurückgegriffen, gelegentlich ergänzt um einen Prätest. Als Indikator für die Feedbackwirkung wird dann die Leistung im Posttest oder die Leistungsentwicklung von Prä- zu Posttest herangezogen. Das Leistungskriterium kann dabei wiederum ein wissens- oder ein verständnisprüfendes Maß sein. Eine eher wissensorientierte Erhebung erfasst den Lernzuwachs oder die Behaltensleistung, wobei im Posttest meist Items der Treatmentphase wiederholt werden. Enthält der Posttest dagegen neue, zu den Items des Treatments mehr oder weniger äquivalente Aufgaben liegt ein eher verständnisorientiertes Maß vor. Die hierbei notwendige Transferleistung kann sich auf die Übertragung bzw. Anwendung von in der Treatmentphase erworbenem deklarativen Wissen (z.B. neue Problemstellungen) oder vermittelten Fähigkeiten oder Prozeduren

(z.B. das Ziehen von Inferenzen, das Anwenden einer erlernten Textstrukturierungsstrategie) beziehen.
Daneben kann zur Beurteilung der Wirksamkeit von Feedback auch die unmittelbare Leistung nach der Gabe der Rückmeldung erfasst und ausgewertet werden. Das heißt, wird Feedback nach falschen Antworten gegeben, interessiert hier, ob dieselbe Aufgabe nach dem Feedback letztlich richtig beantwortet werden konnte oder nicht. Dieses Maß gilt als ein wichtiger Indikator, der jedoch kaum umgesetzt wird (Kulhavy & Stock, 1989). Die Erfassung dieser Korrekturleistung erfordert zudem, dass nach den Rückmeldungen erneute Antwortversuche implementiert sind.
Wenn die Korrekturleistung erfasst wird, liegt es nah, auch die Leistung in den initialen Antworten für Testaufgaben zu erfassen und für die Beurteilung der Wirksamkeit einer Feedbackintervention heranzuziehen. Dieser Indikator spiegelt auch eine Art Transferleistung wider. Der dem zugrunde liegende Gedanke ist der folgende: wenn eine Feedbackart geeignet ist, um aus ihr neue Erkenntnisse zu gewinnen und diese auf neue Aufgabenstellungen transferieren zu können, sollte sich ein Effekt nicht erst in einem anschließenden, konventionellen Transfertest (Posttest ohne Feedback) zeigen, sondern auch schon für ähnliche Aufgaben im Verlauf der Test- bzw. Treatmentphase. Hieran kann auch der dynamische Testgedanke, der für diese Untersuchung relevant ist, angeknüpft werden.

Gelegentlich werden neben den Leistungsmaßen auch die Bearbeitungszeiten bzw. Latenzzeiten herangezogen. Diese dienen allerdings eher der Beurteilung, ob oder wie gewissenhaft Feedback rezipiert wird. Auf der Grundlage der Bearbeitungszeiten wird vereinzelt auch das Kriterium der Effizienz genutzt. Feedbackeffizienz meint nach Kulhavy und Kollegen (1985) das Verhältnis der Menge richtig gelöster Testaufgaben und der für die Rezeption der Feedbacks benötigten Zeit.

Neben dem Leistungskriterium werden Feedbackinterventionen auch hinsichtlich ihrer Auswirkungen auf motivational-emotionale Zustände des Lerners beurteilt. Dafür werden typischerweise Selbsteinschätzskalen im Prä-Post-Vergleich oder *Think aloud*-Prozeduren zur Verbalisierung von Gedanken angewendet.

3.5 Zur Wirksamkeit von Feedback

Die Auseinandersetzung mit den Effekten von Feedback auf Lernen und Leistung ist häufig mit Vorurteilen belegt. Es wird oft davon ausgegangen, *dass* sich Feedback, wenn es angemessen konstruiert ist, positiv auf die Leistung auswirkt, und je mehr Hilfestellungen gegeben werden, desto besser gelingt es, einen Fehler zu korrigieren (Kluger & DeNisi, 1996; Kulhavy & Stock, 1989). Konkret sind es die elaborierten Feedbackarten, denen im Vergleich zu den einfachen Rückmeldungen häufig von vornherein ein größeres Potenzial für die Leistungsverbesserung zugesprochen wird. Diese eher intuitiven Annahmen sind im Wesentlichen auf zwei Aspekte der Feedbackforschung zurückzuführen, die bereits erläutert wurden: 1) Die Intention der Feedbackgabe im pädagogisch-psychologischen Kontext besteht normalerweise in der Förderung von Lernen und Leistung (siehe 3.1). 2) Die Annahmen zur Wirkung von Feedback werden inzwischen vornehmlich vor dem kognitivistischen Paradigma formuliert, wonach sich die Art und der Umfang einer Rückmeldung wesentlich auf den Informationsverarbeitungsprozess und damit die Leistung auswirken sollten (siehe Abschnitt 3.2.1). Vor diesem Hintergrund ist auch nachvollziehbar, dass die Bemühungen der meisten Feedbackstudien auf die Gestaltung des Feedbackinhalts gerichtet sind.

In der Tat kann Feedback ein sehr wirksames Mittel zur Förderung von Lernen und Leistung darstellen (Bangert-Drowns et al., 1991; Hattie, 2009; Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Lysakowski & Walberg, 1982). In der Arbeit von Hattie (2009), in der über 800 Metaanalysen zu über 100 verschiedenen Einflüssen auf die schulische Leistung aggregiert sind, rangiert Feedback mit einem durchschnittlichen Effekt von ES = 0.73 auf dem zehnten Platz der wirksamsten Mittel. Doch spätestens durch die Metaanalysen von Bangert-Drowns und Kollegen (1991) sowie Kluger und DeNisi (1996) wurde ins Bewusstsein gebracht, dass die Effekte von Feedback auf die Leistung zum einen hoch variabel und zum anderen keineswegs immer positiv sind. Beide Metaanalysen zeigen, dass Feedbackinterventionen nicht selten ohne Auswirkungen auf die Leistung bleiben und sogar Verschlechterungen nach sich ziehen können.

Als nachteilig für die Leistung erweist sich in der Metaanalyse von Kluger und DeNisi (1996) Feedback, das aufgabenirrelevante Kognitionen provoziert und damit von der eigentlichen Aufgabenstellung ablenkt, und ohne Wirkung auf die Leistung bleiben Lob und Feedback, welches das Selbstwertgefühl des Lerners bedroht. Bangert-Drowns und

Kollegen (1991) haben überdies herausgearbeitet, dass negative oder ausbleibende Effekte primär aus Studien hervorgehen, die ausschließlich das Feedback Knowledge of Result nutzen und eine Untersuchungsanlage verwenden, bei der der Lerner noch vor dem Antworten die Feedbackmitteilung (ggf. mit richtiger Lösung) einsehen kann[1]. Werden dagegen Knowledge of Correct Result oder Hinweise, die auf das Erreichen der richtigen Antwort ausgerichtet sind, eingesetzt, wirkt sich das im Allgemeinen leistungssteigernd aus. Die durchschnittlichen Effektgrößen hierfür betragen ES = 0.22 und ES = 0.53 (das Mittel aus beiden wird mit ES = 0.31 angegeben). Studien, die entweder Knowledge of Correct Result oder weiterführende Hinweise verwendeten und zusätzlich für die Vorabeinsicht der Rückmeldungen/richtigen Antworten in ihren Interventionen kontrollierten, wirkten sich mit einer mittleren Effektgröße von ES = 0.58 im Vergleich zu ES = 0.31 noch positiver auf die Leistung aus (Bangert-Drowns et al., 1991, S. 228f.).

Auch andere Metaanalysen und Überblicksarbeiten kommen zu dem Schluss, dass Feedback, das über die ausschließliche Gabe von Knowledge of Result hinausgeht, sich positiv auf die Leistung auswirkt. Teilweise wird dabei das Nennen der richtigen Lösung schon als hinreichend eingestuft (Bangert-Drowns et al., 1991; Kluger & DeNisi, 1996), teilweise wird dagegen resümiert, dass darüber hinausgehende Hinweise zur Bewältigung einer Aufgabenstellung, also elaboriertes Feedback, am wirksamsten sind (Hattie, 1999; Hattie & Timperley, 2007). Aus einigen Arbeiten geht allerdings auch nicht eindeutig hervor, welche Art von Feedback tatsächlich gemeint ist, wenn beispielsweise über „Hinweise“ oder „korrigierende Rückmeldungen“ gesprochen wird (z.B. Lysakowski & Walberg, 1982). Eine Einschätzung der Befunde und gegebenenfalls der Empfehlungen fällt deshalb schwer. Dem Feedbackinhalt kommt in jedem Fall eine große Bedeutung für die Effektivität einer Feedbackintervention zu; nach der Kontrolle für die Vorabeinsicht in das Feedback stellt sich der Feedbackinhalt bei Bangert-Drowns und Kollegen (1991)

[1] Dass ein Lerner das Feedback (mit richtiger Lösung) einsehen kann, noch bevor er eine Antwort auf eine Aufgabenstellung abgibt, ist in der englischsprachigen Literatur mit dem Begriff „*presearch availability*“, eingeführt von Kulhavy (1977, S. 217), belegt. Dieses Merkmal einer Lernumgebung ist auf die eingeschränkten (technischen) Möglichkeiten früherer Untersuchungen zurückzuführen. Es kann dazu führen, dass die Probanden statt das eigentliche Lernmaterial zu bearbeiten und selbstständig nachzudenken lediglich den Inhalt der Rückmeldung abschreiben/nutzen, um die Aufgabe zu beantworten. Auf diese Art findet kein Lernen statt und eine Leistungssteigerung bei der Überprüfung des Treatments (Posttest) wird unwahrscheinlich.

als zweitwichtigste Einflussgröße für die Feedbackwirksamkeit heraus. Beide Variablen zusammen erklärten etwa die Hälfte der Varianz in den Effektgrößen.

Mit Bezug auf die erwähnten Vorurteile bezüglich der Feedbackwirksamkeit ist im Allgemeinen also festzuhalten, dass Feedback wirksam sein kann, es aber nicht automatisch ist. Der Art der rückgemeldeten Information kommt tatsächlich eine wichtige Stellung zu. Dass jedoch mehr und spezifischere Informationen besser sind als weniger, kann nicht allgemeingültig bestätigt werden. Dabei erweisen sich vor allem zwei Aspekte der Metaanalysen und Überblicksarbeiten zur Feedbackwirksamkeit als problematisch: Zum einen findet über die Bezeichnung „weiterführende Hinweise“ hinaus keine Differenzierung statt. Zu den elaborierten Feedbackarten zählen aber nicht nur Hinweise, die beispielsweise auf die für die Lösung einer Aufgabe relevanten Prozesse/Schritte gerichtet sind, sondern sie könnten sich auch auf relevante Textstellen beziehen und dergleichen mehr (vgl. Tabelle 2, S. 35). Zum anderen erweist es sich als problematisch, dass in den Studien keine Differenzierung hinsichtlich der Test- bzw. Aufgabenanforderungen, für die Feedback gegeben wird und für die der Lerner profitieren soll, stattfindet.

Die Art der Aufgabe (z.B. Gedächtnis- vs. Transferaufgabe) bzw. die Aufgabenanforderungen wie Komplexität stellen ein wichtiges Kriterium für die Feedbackwirksamkeit dar (Kluger & DeNisi, 1996). Die Befunde der Metaanalyse von Bangert-Drowns und Kollegen (1991), die vergleichsweise am stärksten hinsichtlich der Feedbackinhalte differenzieren, beziehen sich aber hauptsächlich auf Studien mit Wissens- oder Gedächtnisaufgaben. Die Übertragbarkeit auf die hierarchiehöheren Anforderungen des Textverstehens, wie sie auch Gegenstand der vorliegenden Arbeit sind, ist nicht ohne weiteres gewährt (Bangert-Drowns et al., 1991; Winne et al., 1993). Für diese höheren Aufgabenanforderungen wie das Ziehen von Inferenzen oder das Anwenden von Regeln in komplexen Lernumgebungen wird vermutet, dass sie den Einsatz von elaboriertem Feedback notwendig machen (Bangert-Drowns et al., 1991). Eine unmittelbare Bestätigung durch die Metaanalysen und Überblicksarbeiten ist aus den genannten Gründen nicht möglich und so erweisen sich deren Befunde zu den Effekten von Feedback auf die Leistung für die vorliegende Arbeit mit ihrem Schwerpunkt auf dem Bereich des Textverstehens als nicht hinreichend informativ.

Deshalb wurden Studien zusammengetragen, die Aussagen über die Wirksamkeit von Feedbackarten (Feedbackinhalten) für das *Textverstehen* ermöglichen. Sie werden im Folgenden dargestellt. Die Anzahl dieser Studien ist allerdings stark begrenzt. Zur

Unterstützung der Argumentation werden (deshalb) auch Anleihen aus dem verwandten Forschungsbereich zum *Prompting* gemacht. Anschließend werden Befunde zu Zeitpunkt, Präsentationsmodus und die Feedbackwirksamkeit beeinflussende Merkmale des Lerners skizziert.

3.5.1 Faktor Feedbackgestaltung

3.5.1.1 Zur Wirksamkeit von Feedbackinhalten

Einige der Feedbackstudien im Bereich des Textverstehens haben sich (auch) den einfachen Feedbackarten gewidmet. Doch weder für Knowledge of Correct Result (Lee, Lim & Grabowski, 2009; Morrison et al., 1995; Peverly & Wood, 2001) noch Answer-until-Correct Feedback (Morrison et al., 1995) konnten Effekte auf das Textverständnis bzw. die Lesekompetenz nachgewiesen werden. Für das am wenigsten umfangreiche Feedback Knowledge of Result sprechen die Befunde ebenfalls dafür, dass es keinen Effekt auf die Leistung hat (Lee et al., 2009). Auch in Kombination mit der zusätzlichen Vermittlung von Lesestrategien, wie von Schunk und Rice (1991, 1993) durchgeführt, konnte über einen möglichen Übungseffekt hinaus kein Effekt von Knowledge of Result auf die Lesekompetenz nachgewiesen werden.

Die Studie von Peverly und Wood (2001) zum Knowledge of Correct Result ist an dieser Stelle noch einmal hervorzuheben. Denn das Feedback wurde hier in einer interessanten Erweiterung verwendet: nach der Beantwortung der Testfragen erhielten die Probanden auf Papier die richtigen Lösungen zu den Fragen. Diesen Feedbackbogen sollten sie sich nicht nur durchlesen, sondern sie waren auch angehalten, ihre eigenen, ursprünglich falschen Antworten anhand des Feedbacks, also der richtigen Lösungen, schriftlich zu korrigieren[2]. Damit wollten die Autoren sicherstellen, dass das Feedback tatsächlich rezipiert wird – was in jeder Feedbackintervention einen kritischen Punkt darstellt. Nichtsdestotrotz sind die Ergebnisse der Studie dahingehend zu bewerten, dass diese Form der Rückmeldung keine generelle Leistungssteigerung nach sich ziehen kann.

[2] Diese Variante der Rückmeldung wird auch *Forced Knowledge of Result* genannt (z.B. Dempsey, Driscoll & Litchfield, 1993).

Zunächst ist also festzuhalten, dass die einfachen Feedbackarten im Allgemeinen nicht das Potential haben, das Textverständnis bzw. die Lesekompetenz zu verbessern. Eine Erklärung hierfür ist, dass Knowledge of Result dem Lerner zwar signalisiert, dass ein Fehler gemacht wurde, aber „it does not provide any information that would further the learner's knowledge or provide additional insight into possible errors in understanding" (Mason & Bruning, 2001, Abs. 8). Zwar liegen Hinweise vor, dass initiiert durch das Fehlersignal unter bestimmten Umständen Lerner selbstständig/selbstreguliert entsprechende Such- und Abrufprozesse anstrengen und erfolgreich im Sinne der korrekten Lösung einer Aufgabenstellung umsetzen können (Song & Keller, 2001; vgl. auch Abschnitt 3.3 Funktionen von Feedback). Doch für den Bereich des Textverstehen bzw. der Lesekompetenz gilt dies im Allgemeinen offensichtlich nicht, was daraufhin deutet, dass hier Fehler, sofern es sich nicht um Flüchtigkeitsfehler handelt, eher Ausdruck von Defiziten in Fähigkeiten der Textverarbeitung oder im Vorwissen sind (vgl. auch Abschnitt 2.3). Beide Ursachen lassen sich kaum spontan und nur mit der Rückmeldung, dass ein Fehler gemacht wurde, lösen. Hier werden eher konkrete Anleitungen, etwa die Vermittlung von Strategien, benötigt. Dass das einfache Feedback Knowledge of Result aber auch dann nicht effektiv ist, wenn wie bei Schunk und Rice (1991, 1993) am Anfang des Treatments Lesestrategien vermittelt wurden, erscheint dann erklärungsbedürftig. Zwar lässt sich dieser Befund aufgrund der Untersuchungsdesigns und den in den Arbeiten angebotenen Informationen nicht mit Sicherheit klären. Aber unter Berücksichtigung des gesamten Untersuchungskontextes (siehe unten, zu elaborierten Feedbackarten) ergeben sich Zweifel, ob die Strategievermittlung am Anfang des Treatments überhaupt erfolgreich war, in dem Sinne, dass die Leser die Strategien selbstreguliert anwenden konnten. In diesem Fall hätten die entsprechenden Probanden auch nicht unter einer Bedingung aus Strategietraining plus Knowledge of Result gearbeitet, sondern sie hätten nur das Feedback erhalten und die Ineffektivität der Intervention wäre damit nachvollziehbar.

Die allgemeinen Erläuterungen zur Ineffektivität des Feedbacks Knowledge of Result gelten in ähnlicher Weise auch für Answer-until-Correct, denn letzeres ist nicht mehr als das Wiederholen von Knowledge of Result bis zur richtigen Lösung. Dass letztlich die richtige Antwort bekannt wird, stellt den einzigen, den wesentlichen inhaltlichen Mehrwert von Answer-until-Correct gegenüber Knowledge of Result dar. Aber auch das Kennen der richtigen Antwort erweist sich als ungeeignet, um das Textverständnis bzw.

die Lesekompetenz zu fördern, wie auch die Befunde zum Knowledge of Correct Result zeigen. Während Knowledge of Correct Result hinreichend informativ ist, um Textinformationen später effektiv erinnern und abrufen zu können (Bangert-Drowns et al., 1991; Lysakowski & Walberg, 1982), reicht sie für die Anforderungen des Textverstehens nicht mehr aus. Das heißt, das Kennen der richtigen Antwort einer (Reihe von) Aufgabe(n) kann für spätere Anforderungen des Textverstehens nicht nutzbar gemacht werden. Für das Verständnis der aktuellen Aufgabenstellung kann Knowledge of Correct Result noch hilfreich sein, etwa um selbstständig zu extrahieren, warum die abgegebene Antwort falsch ist. Doch damit Lerner bei folgenden Aufgaben von dieser Art der Rückmeldung profitieren könnten, setzt das voraus, dass aus den rückgemeldeten richtigen Lösungen ein Prinzip geschlussfolgert und dieses auf andere, ähnliche Aufgabenstellungen richtig transferiert wird. Diese Anforderung wird allem Anschein nach von Lernern im Allgemeinen nicht geleistet. Insgesamt spricht das für die Vermutung von Bangert-Drowns und Kollegen (1991), die formulieren, dass für höhere Aufgabenanforderungen wie das Ziehen von Inferenzen anstatt einfacher Rückmeldungen eher elaborierte Feedbacks relevant seien.

Die Befunde zum Nutzen elaborierter Feedbackarten im Bereich des Textverstehens ergeben ein gemischtes Bild. Sie können zu einer Verbesserung führen (Lee et al., 2009; Schunk & Rice, 1991, 1993; van den Boom, Paas & van Merrienboer, 2007; Winne et al., 1993), einige bleiben aber auch ohne Wirkung auf das Textverständnis bzw. die Lesekompetenz (Schunk & Rice, 1986, 1993; Winne et al., 1993). Da sich hinter der Bezeichnung „elaboriertes Feedback“ per se verschiedene Arten verbergen, die zudem in den jeweiligen Untersuchungskontexten unterschiedlich ausgestaltet sind, werden die Studien im Folgenden näher erläutert.

In der Untersuchung von Winne und Kollegen (1993) wurde ein Interventionsprogramm zum Textverstehen umgesetzt, in dessen Rahmen die teilnehmenden Schüler kürzere Texte rezipierten und Verständnisfragen beantworteten. Die Verständnisfragen erforderten entweder das Ziehen einer textbasierten Inferenz oder das Auffinden explizit genannter Information. Für die Antworten erhielten die Schüler von einem Tutor Feedback: In der ersten Treatmentbedingung, dem „*inductive feedback*“, wird zunächst die Korrektheit der Antwort mitgeteilt, im Falle einer falschen Antwort wird dann noch die richtige Lösung genannt, und der Tutor markiert im Text die zur Lösung der Aufgabe relevanten Informationen. In der zweiten Treatmentbedingung, dem „*explicit feedback*“,

gibt der Tutor dieselbe Rückmeldung wie in der ersten Bedingung, erklärt darüber hinaus aber noch, warum die markierten Informationen relevant und andere Textstellen wiederum irrelevant für die Beantwortung der jeweiligen Aufgabe sind. Bei den Aufgaben zum Ziehen textbasierter Inferenzen erläutert er zudem, wie die Inferenz zu ziehen ist. Die Analysen zeigen einen Effekt allein für die Feedbackvariante „explicit feedback“, und dies wiederum nur für die Aufgaben des Typs Inferenzen Ziehen: Schüler, denen im Training „explicit feedback“ präsentiert wurde, beantworteten im Posttest mehr inferenzielle Fragen richtig (Cohens d = .86) als Schüler der Bedingung „inductive feedback“ und sie konnten zudem im Gegensatz zur Vergleichsgruppe einen Leistungszuwachs vom Prä- zum Posttest auf statistisch signifikanten Niveau verzeichnen (Cohens d = .60). Diese Befunde können also dahingehend interpretiert werden, dass die Erklärung, warum bestimmte Informationen relevant sind und wie daraus die gewünschte Inferenz zu ziehen ist, den Mehrwert der Feedbackintervention ausmachen.

Schunk und Rice (1986, 1991, 1993) untersuchten in einer Reihe von Experimenten die Wirksamkeit verschiedener elaborierter Feedbacks in Kombination mit der Vermittlung einer fünfstufigen Lesestrategie. Diese bestand aus den folgenden Handlungsanweisungen für den Umgang mit Texten: „What do I have to do? 1.) Read the questions. 2.) Read the passage to find out what it is mostly about. 3.) Think about what the details have in common. 4.) Think about what would make a good title. 5.) Reread the story if I don’t know the answer to a question.” (z.B. Schunk & Rice, 1986, S. 59). Die Lesestrategien wurden jeweils am Anfang der Interventionen vermittelt, dabei mit den Schülern besprochen, und im Verlauf der Interventionssitzungen wiederholt aufgegriffen. Die Schüler lasen während der Treatmentsitzungen verschiedene Texte mit ansteigender Länge (4 bis hin zu 25 Sätze lang in Schunk & Rice, 1986, 1991; keine genaue Angaben in Schunk & Rice, 1993) und beantworteten Multiple-Choice Items, die das Erfassen der Hauptaussagen eines Textes erforderten. Die fünfstufige Lesestrategie blieb in allen drei Studien während des gesamten Treatments sichtbar an einer Tafel stehen und die Lehrerin erinnerte die Schüler gelegentlich daran, diese wie instruiert anzuwenden. Für die Antworten auf die Textfragen bekamen die Probanden Feedback, das durch die Lehrerin verbal und, trotz der Arbeit in Kleingruppen, jedem einzeln vermittelt wurde.
In Schunk und Rice (1986) wurde das Feedback ausschließlich nach richtigen Antworten gegeben. Die Rückmeldungen bestehen in einer Zuschreibung des Erfolgs entweder auf die Fähigkeit (z.B. „You‘re good at this.“) oder die Anstrengung des Schülers (z.B.

„You've been working hard."). Die Ergebnisse zeigen, dass keine Variante dieses elaborierten Feedbacks, weder die Anstrengungs- noch die Fähigkeitszuschreibung von Erfolg, einen Leistungszuwachs im Textverstehen nach sich ziehen konnte.

In Schunk und Rice (1991) wurde einfaches mit elaboriertem Feedback verglichen. Alle Probanden erhielten für ihre Antworten Knowledge of Result (z.B. „That's correct.") und die Gruppe mit dem elaborierten Feedback (im Original „*progress feedback*" genannt) erhielt zusätzlich eine Rückmeldung, die den individuellen Fortschritt im Anwenden der Strategie für die Beantwortung der Items zum gelesenen Text rückmeldet. Beispiele für diese Art des Feedbacks lauten: „That's correct. Your're learning to use the steps." oder „That's correct. You got it right because you followed the steps in order." (Schunk & Rice, 1991, S. 358). Das elaborierte Feedback wurde allerdings nicht immer, sondern pro Sitzung drei- bis viermal, und offenbar tatsächlich nur bei richtigen Antworten gegeben. Hierzu werden zwar keine genauen Angaben gemacht, aber diese Variante erscheint vor dem Hintergrund, dass sich die elaborierten Rückmeldungen auf den Fortschritt der Schüler beziehen, plausibel.

Die Befunde zeigen, dass unter Kontrolle der Leistung im Prätest Schüler aus der Bedingung des elaborierten Feedbacks signifikant mehr Aufgaben zum Textverstehen im Posttest richtig lösten als Schüler, die ausschließlich Knowledge of Result erhielten. Die Wirksamkeit des elaborierten Feedbacks scheint allerdings eher darin zu liegen, dass es an die Anwendung der Strategie erinnert statt tatsächlich über die individuelle Entwicklung bei der Strategienutzung zu informieren. Die im Original angeführten Beispiele des Feedbackinhalts legen diesen Schluss zumindest nah.

In einer späteren Untersuchung verglichen Schunk und Rice (1993) wiederum einfaches mit elaboriertem Feedback und setzten dabei die gleichen Arten von Rückmeldungen wie in Schunk und Rice (1991) ein, allerdings weicht die Bezeichnung des elaborierten Feedbacks ab. Wieder erhielten alle Probanden Knowledge of Result für ihre Antworten und den Probanden in der Gruppe des elaborierten Feedbacks wurde für drei bis vier richtige Antworten pro Sitzung zusätzlich die richtige Anwendung der Strategie rückgemeldet (im Original „strategy value feedback" genannt, z.B. „You got it right because you followed the steps in the right order." oder "Do you see how thinking about what the details have in common helps you answer questions?", S. 266). Daneben wurde in dieser Studie, anders als in Schunk und Rice (1991) und Schunk und Rice (1986), als zweiter Faktor die Art der Strategievermittlung bzw. –anwendung variiert. Die Hälfte der Stichprobe wurde dahingehend trainiert, die fünfstufige Lesestrategie im Verlauf der

mehrteiligen Treatmentphase zu internalisieren. Von der anderen Hälfte der Stichprobe wurden die Strategien immer wieder laut aufgesagt.

Die Ergebnisse zeigen, dass das elaborierte Feedback zu einer höheren Leistung im Posttest führte als Knowledge of Result, wenn die Lesestrategie nicht internalisiert war. (Damit ist auch das Ergebnis der Vorgängerstudie der Autoren von 1991 repliziert.) In den beiden Feedbackgruppen, die die Lesestrategie dagegen internalisiert hatten, lag kein Leistungsunterschied im Posttest vor. Zudem zeigen die deskriptiven Statistiken in der Arbeit, dass die durchschnittlichen Leistungen beider Feedbackgruppen mit Strategieinternalisierung deutlich über den jeweiligen Feedbackbedingungen ohne Internalisierung der Strategie liegen. Eine entsprechende Analyse des möglichen Haupteffekts der Art der Strategievermittlung auf die Leistung ist jedoch nicht aufgeführt. Aber anhand der berichteten Mittelwerte und Standardabweichungen lassen sich die Effektgrößen (Cohens *d*) berechnen und die betragen für Knowledge of Result $d = 2.12$ und für das elaborierte Feedback $d = 1.43$, jeweils zugunsten der Bedingung der Strategieinternalisierung gegenüber keiner Internalisierung. Das heißt also, bei einer Strategieinternalisierung wird die Posttestleistung im Vergleich zu einer Testbearbeitung ohne Internalisierung angehoben, die Art des gegebenen Feedbacks in der Intervention bewirkt aber keinen Leistungsunterschied. Ohne Strategieinternalisierung gibt es jedoch einen Effekt der Feedbackart auf die Posttestleistung, und zwar zugunsten des elaborierten Feedbacks.

Die Ergebnisse werden dahingehend interpretiert, dass bei einer nicht vorhandenen Strategieinternalisierung der Einsatz eines elaborierten Feedbacks, dass auf die Anwendung der Strategie(n) verweist, leistungsförderlich ist. Ohne diese Hinweise, also bei einer Feedbackintervention nur mit Knowledge of Result, ist eine Leistungssteigerung über einen Übungseffekt hinaus fraglich, vielleicht weil die Strategien eben nicht angewendet werden. Die Beherrschung der Lesestrategie(n) wirkt sich deutlich leistungssteigernd aus und hier bringt eine von außen angeregte Anwendung ebendieser Strategien durch entsprechende Hinweise auch keinen Vorteil mehr, zumindest in einem Testsetting wie in der Studie von Schunk und Rice (1993).

Eine weitere Umsetzung von elaboriertem Feedback im Kontrast zur alleinigen Darbietung von Knowledge of Correct Result findet sich in der Studie von Lee und Kollegen (2009), die eine computerbasierte Lernumgebung einsetzten. Obwohl für die Autoren die Nützlichkeit von Feedbacks für den Wissenserwerb im Mittelpunkt stand,

sind aufgrund der Auswahl der Tests auch Aussagen zum Textverstehen möglich. Das Ziel bei der Bearbeitung der Lernumgebung bestand darin, sich Wissen zu einem komplexen Sachthema anzueignen. Dazu wurden mehrere Textpassagen und dazugehörige Testfragen präsentiert. Es wurden eine Kontroll- und zwei Treatmentbedingungen kontrastiert. Beide Treatmentbedingungen erhielten im Verlauf des Lernprogramms explizite Aufforderungen, einzelne Strategien wie das Markieren wichtiger Informationen oder das Zusammenfassen des eigenen Verständnisses anzuwenden. Außerdem bekamen sie für ihre Antworten auf die Testfragen Feedback: entweder Knowledge of Correct Result (Markierung der richtigen Antwortalternative) oder ein elaboriertes Feedback, das zusätzlich zum Knowledge of Correct Result noch einen metakognitiven Hinweis präsentierte, der sich sehr wahrscheinlich auf die aktive Strategieanwendung bezog. Das genaue Prinzip wurde nicht genannt (vgl. Lee et al., 2009, S. 21), doch das in der Originalarbeit angegebene Beispiel einer Rückmeldung legt diese Vermutung nah: „Incorrect! You need to go back and revise your highlighted sentence or summary." (S. 14) Das Besondere der beiden Treatmentbedingungen ist, dass die im Lernprogramm eingeflochtenen Aufforderungen zur Strategieanwendung selbst schon eine Hilfestellung darstellen, nur dass diese Hilfe eben nicht an eine konkrete Antwort gebunden ist. Beide Treatmentbedingungen werden schließlich noch mit einer Kontrollbedingung kontrastiert, deren Probanden weder die Hinweise zur Strategieanwendung bei der Bearbeitung der Texte noch Feedback für ihre Antworten auf die Testfragen erhielten.

Die Ergebnisse zeigen, dass die Versuchsbedingung mit dem elaborierten Feedback eine signifikant höhere Leistung hervorbrachte als die Kontrollbedingung. Die Leistung in der Versuchsbedingung mit dem einfachen Feedback Knowledge of Correct Result liegt dazwischen: der Unterschied zur Kontrollbedingung einerseits und zum elaborierten Feedback andererseits ist nicht statistisch bedeutsam. Auch wenn die Autoren der Studie keine Erklärung für dieses Ergebnismuster liefern, kann vermutet werden, dass durch die Prompts zur Strategieanwendung beide Feedbackgruppen profitieren, die mit dem elaborierten (metakognitiven) Feedback aber durch den expliziten Anwendungsbezug zur Strategie den Mehrwert beinhaltet, damit die Leistung im Vergleich zur Kontrollbedingung (ohne Strategieprompt, ohne Feedback) auch statistisch bedeutsam wird.

Van den Boom und Kollegen (2007) kombinierten ebenfalls die Vermittlung einer metakognitiven Strategie mit elaboriertem Feedback, das die Anwendung der Strategie anregt. Im Rahmen einer computerbasierten Lernumgebung wurden Studenten angehalten, über ihren Lernprozess zu reflektieren und dies schriftlich festzuhalten. Ein Teil der Stichprobe erhielt zusätzlich elaboriertes Feedback, das zur Anwendung dieser reflexiven Strategie auffordert, indem es indirekte Hinweise auf Fehler oder Probleme bei den Reflexionen bietet. Die Autoren bezeichnen diese Form als suggestives Feedback. Gemessen anhand eines Tests zu den Inhalten des Kurses, der mit großer Wahrscheinlich eine verständnisorientierte Überprüfung anstellt (genaue Angaben werden hier allerdings nicht gemacht), zeigt sich, dass Studenten, die elaboriertes Feedback erhielten eine signifikant bessere Testleitung erbrachten als Studenten, die ohne Feedback bezüglich ihrer Reflexionen über das Lernmaterial arbeiteten, und auch besser als Studenten einer Kontrollbedingung (keine Reflexionen, kein Feedback).

Am Rande sei noch die Studie von Murphy (2010) erwähnt, auch wenn sie sich zugegebenermaßen nicht nahtlos an die vorherigen Feedbackstudien anfügt. Sie behandelt zwar auch die Wirksamkeit von Feedback für Textverstehen, jedoch im Bereich des Zweitspracherwerbs und die Untersuchungsteilnehmer arbeiten hier in Zweiergruppen an einer computerbasierten Lernumgebung, in deren Rahmen Aufgaben zum Textverstehen zu beantworten waren. Während der Bearbeitung der Fragen erhielten die Zweiergruppen je nach Versuchsbedingung eine Form von Feedback: entweder Knowledge of Correct Result oder eine Rückmeldung, die vor Knowledge of Correct Result zunächst elaboriertes Feedback in Form von Hinweisen, die die Interaktion fördern und die Zweiergruppen in ihren Bemühungen, fehlerhafte Antworten selbst zu korrigieren, anregten. Die Ergebnisse belegen, dass das elaborierte Feedback zu einem substanziell besseren Textverständnis führte als Knowledge of Correct Result allein. Zwar lässt sich ein gewisser Außenseitercharakter dieser Studie nicht abstreiten. Dennoch könnte sie die bisherigen Darstellungen dahingehend bekräftigen, dass Rückmeldungen, die eine Auseinandersetzung mit Merkmalen einer Aufgabe, mit fehlerhaften Antworten und möglicherweise mit Lösungsansätzen anregen, förderlich für das Verständnis eines gelesenen Textes sein können.

In Tabelle 4 sind die wichtigsten Merkmale und die zentralen Aussagen der erläuterten Studien zusammengefasst.

Tabelle 4 Zusammenfassung der Feedbackstudien zum Textverstehen

Quelle	Merkmale des Designs	Versuchsbedingungen (Feedbackarten)[a]	Effekte[b]
Morrison, Ross, Gopala-krishnan & Casey (1995)	Feedbackquelle: Computer, Schriftliches Feedback, Einmalige Treatmentsitzung, N = 246 Studenten	a. Answer-until-Correct, b. Knowledge of Correct Result, c. verzögertes Knowledge of Correct Result, d. Kein Feedback (Kontrolle 1), e. Keine Fragen zum Text (Kontrolle 2)	Kein Effekt für Verständnisfragen
Peverly & Wood (2001)	Feedbackquelle: Lernumgebung, Schriftliches Feedback, Mehrere Treatmentsitzungen, N = 50 Schüler (14-16 Jahre alt) mit Lernschwierigkeiten	a. Knowledge of Correct Result mit „erzwungener" Antwortkorrektur, b. Kein Feedback (Kontrolle)	kein Haupteffekt
Winne, Graham & Prock (1993)	Feedbackquelle: Tutor, Mündliches und schriftliches Feedback, Mehrere Treatmentsitzungen, Einzelsitzungen, N = 24 Schüler (3.-5.Klasse), schlechte Leser mit Lernschwierigkeiten	Immer Knowledge of Correct Result bei Fehlern/Knowledge of Result bei richtigen Antworten plus: a. Inductive Feedback (Tutor markiert relevante Informationen im Text), b. Explicit Feedback (Tutor erklärt zusätzlich zu Markierungen, warum diese relevant sind und wie Inferenzen zu ziehen sind)	a<b für Inferenzfragen im Posttest
Schunk & Rice (1986)	Feedbackquelle: Person, Mündliches Feedback, mehrere Treatmentsitzungen, Arbeit in Kleingruppen mit je einer Leiterin, N = 40 Schüler (M=10;8 Jahre), mit Problemen im Textverstehen	Immer Knowledge of Result plus: a. Attribuierung des Erfolgs auf Fähigkeit, b. Attribuierung des Erfolgs auf Anstrengung	Kein Zuwachs von Prä- zu Posttest
Schunk & Rice (1991)	Siehe Schunk und Rice (1986), N = 30 Schüler (M=11;3 Jahre)	a. Knowledge of Result, b. Progress Feedback (bezogen auf Erfolg im Anwenden der Lesestrategie)	a<b im Posttest
Schunk & Rice (1993)	Siehe Schunk und Rice (1986), N = 44 Schüler (M=10;8 Jahre)	1. Faktor: Lesestrategie internalisiert/ nicht internalisiert, 2. Faktor: Feedback: a. Knowledge of Result, b. Strategy Value Feedback (Erfolg wird auf Anwendung der Strategie bezogen)	Lesestrategie internalisiert: Kein Effekt im Posttest Lesestrategie nicht internalisiert: a<b
Lee, Lim & Grabowski (2009)	Feedbackquelle: Computer, Schriftliches Feedback, Mehrere Treatmentsitzungen, N = 36 Studenten	a. Knowledge of Correct Result, b. Knowledge of Correct Result plus metakognitives Feedback (Verweis auf Strategieanwendung), c. Kein Feedback (Kontrolle)	b>c; a ohne signifikanten Effekt zu b oder c

Noch Tabelle 4 Zusammenfassung der Feedbackstudien zum Textverstehen

Quelle	Merkmale des Designs	Versuchsbedingungen (Feedbackarten)[a]	Effekte[b]
Van den Boom, Paas & van Merrienboer (2007)	Feedbackquelle: Tutor, aber vermittelt über Computer, Schriftliches Feedback, Mehrere Treatmentsitzungen, N = 49 Studenten	a. Instruktion zu Reflektieren mit elaboriertem Feedback (weist indirekt auf Probleme/Fehler in Reflexionen zum Lernstoff hin, z.B. „Can you explain why you did it that way?“), b. Instruktion zu Reflektieren, kein Feedback (Kontrolle 1), c. keine Instruktion zu Reflektieren, kein Feedback (Kontrolle 2)	a>b, c im Posttest
Murphy (2010)	Feedbackquelle: Lern-umgebung, Schriftliches Feedback, Einmalige Treatmentsitzung, Immer 2 Schüler arbeiten zusammen, N = 267 Studenten	a. Knowledge of Correct Result, b. Elaboriertes Feedback (unterstützt Zusammenarbeit um Fehler selbst/gegenseitig zu korrigieren), gefolgt von Knowledge of Correct Response	a<b

Anmerkungen. [a] Einige Studien haben neben den Feedbackbedingungen einen zweiten Faktor untersucht. Sofern dieser in den interessierenden abhängigen Variablen keine Interaktionseffekte mit Feedback aufweist, wird er hier nicht berichtet.
[b] Die dargestellten Effekte beziehen sich nur auf die Maße, die Aussagen über Textverstehen/Lesekompetenz erlauben.

Integration der Befunde und Bewertung

Um Merkmale potentiell erfolgreicher Rückmeldungen für den Bereich des Textverstehens/der Lesekompetenz benennen zu können, werden nachfolgend Merkmale der erfolgreichen und der nicht erfolgreichen Feedbackinterventionen diskutiert und in Bezug zu allgemeineren Erkenntnissen der Feedbackforschung gesetzt.

Die meisten der erläuterten elaborierten Feedbackarten haben zu einer Verbesserung des Textverständnisses bzw. der Lesekompetenz geführt. Der Vergleich dieser erfolgreichen Rückmeldungen lässt im Wesentlichen auf zwei inhaltliche Ansätze schließen:

- Feedbacks beziehen sich auf die Anwendung von (zuvor vermittelten) kognitiven oder metakognitiven Lesestrategien.
- Feedbacks fokussieren die für die Bewältigung einer Aufgabenstellung relevanten Teilfähigkeiten/-prozesse des Textverstehens; bei Winne und Kollegen (1993) betrifft es das Ziehen von Inferenzen, das durch Erklärungen, warum bestimmte Textinformationen relevant sind und wie die Inferenzen zu ziehen sind, unterstützt wird.

In der Tat beinhalten die für den Bereich des Textverstehens dargestellten Feedbackstudien sehr häufig eine Strategievermittlung. Die Strategien werden entweder am Anfang eines Treatments instruiert und eingeübt (z.B. in Schunk & Rice, 1986, 1991, 1993) oder ohne explizite Instruktion innerhalb einer Lernumgebung als Hinweise eingestreut (z.B. in Lee et al., 2009). Lesestrategien an sich stellen ein wirksames Mittel zur Unterstützung des Leseverständnisses dar (vgl. Abschnitt 2.4). Sie können als Pläne für Handlungsabfolgen (Klauer, 1988) bzw. als konkrete Techniken beschrieben werden, die das Verstehen und Behalten schriftlicher Informationen unterstützen (Artelt, 2000). Ihre Einbindung in feedbackgestützte Interventionen erscheint daher plausibel und naheliegend. Vor diesem Hintergrund ist auch das Ergebnis aus Schunk und Rice (1993) zu erklären, wonach die feedbackvermittelten Hinweise zum Anwenden der Lesestrategien dann redundant zu werden scheinen und keinen Nutzen mehr zeigen, wenn die Leser schon gute Strategienutzer sind, das heißt die Lesestrategie (wie in dieser Studie geschehen) internalisiert haben und damit sehr wahrscheinlich selbstständig und spontan Strategien zum Verständnis von Texten einsetzen. Auf der anderen Seite kostet eine dem Treatment vorangestellte Strategievermittlung bzw. ein Strategietraining zusätzliche Zeit und das Anregen der Strategieanwendung durch Feedback setzt voraus, dass die Strategien prinzipiell beherrscht werden und ausführbar sind.

Dass auch Rückmeldungen effektiv für das Textverständnis bzw. Textverstehen sind, die wie bei Winne und Kollegen (1993) über die für die Bewältigung einer vorliegenden Aufgabenstellung relevanten Teilprozesse informieren, lässt sich gut an die Erläuterungen zum Nutzen von Lesestrategien in/mit Feedbacks anknüpfen. Die Lesestrategie stellt eher eine allgemeine Handlungsabfolge oder „Schablone“ dar (z.B. Verbindungen im Text suchen). Rückmeldungen wie bei Winne und Kollegen (1993) sind in dieser Hinsicht spezifischer, weil sie konkret – also für die jeweilige Aufgabenstellung bzw. die entsprechende Textpassage – vorgeben, was zu tun ist, also beispielsweise, welche Informationen in dem entsprechenden Textabschnitt wie zu verbinden sind.
Spezifische Rückmeldungen gelten im Allgemeinen als effektiv und werden daher für die Gestaltung von Rückmeldungen empfohlen (Balzer, Doherty & O'Connor, 1989; Shute, 2008). Allerdings wird diese generelle Einschätzung an anderer Stelle (Brehmer, 1979; Hattie & Timperley, 2007; Thompson, 1998) relativiert: (sehr) spezifisches Feedback scheint vor allem für die Aufgaben zu nützen, für die es gegeben wird. Die Übertragung

auf andere Aufgaben (z.B. in einem Posttest) ist dagegen häufig eingeschränkt. Da die oben berichteten Befunde allerdings ausnahmslos auf Transferleistungen beruhen und dabei eben positive Auswirkungen auf die Leistung zu beobachten waren, spricht das für die Nutzbarkeit der berichteten Rückmeldungen. Es ist zu vermuten, dass aus den konkreten Rückmeldungen bezüglich des Generierens von Inferenzen generelle Prinzipien, Strategien, erlernt wurden. Dabei ist allerdings auch zu berücksichtigen, dass die Studie von Winne und Kollegen (1993) auf mehrmaligen Treatmentsitzungen, also ausgedehnteren Lernphasen beruht.

In den berichteten Feedbackstudien zum Textverstehen fand ein inhaltlicher Aspekt, der ansonsten häufig zusammen mit der Spezifität von Rückmeldungen thematisiert wird, keine Beachtung: die Komplexität. Auf der Grundlage der allgemeinen Feedbackliteratur ist festzuhalten, dass Untersuchungen zum Einfluss der Komplexität von Rückmeldungen auf ihre Wirksamkeit gegenläufige Befunde erbrachten. Einerseits sind Zusammenhänge zwischen der Komplexität von Feedback und seinen Effekte feststellbar, und zwar negative (Kulhavy et al., 1985). Andererseits können keine Einflüsse der Feedbackkomplexität nachgewiesen werden (vgl. Shute, 2008). Shute (2008) resümiert, dass die Wirksamkeit von Feedback wohl nicht einfach von der Menge an vermittelter Information abhängt, sondern dass anderen Faktoren eine bedeutendere Rolle zukommt. In erster Linie wird hier die bereits erläuterte Spezifität des Feedbackinhalts gesehen.

Weitere Gemeinsamkeiten der erläuterten Feedbackstudien aus dem Bereich des Textverstehens sind, dass die Rückmeldungen:

- meist nach Fehlern gegeben werden und
- fast immer auf der kriterienbezogenen Bezugsnorm, d.h. auf einem objektiven Lern- oder Leistungskriterium, beruhen (soweit das aufgrund der Ausführungen in den Veröffentlichungen der Studien beurteilt werden kann).

Bei Winne und Kollegen (1993) fließt zwar gelegentlich auch die intraindividuelle bzw. persönliche Ebene ein, indem Lerner nach richtigen Antworten auch gelobt werden. Doch die elaborierten Rückmeldungen nach Fehlern, denen im Gegensatz zu richtigen Antworten mehr Gewicht für den Lernprozess beigemessen wird (Kulhavy, 1977), konzentrieren sich wiederum auf das Leistungskriterium. Allein in der Untersuchungsreihe von Schunk und Rice (1986, 1991, 1993) weisen die Rückmeldungen eine starke Betonung der intraindividuellen Ebene auf (z.B. „You've been answering a lot

more questions correctly since you've been using these steps.", aus Schunk & Rice, 1993, S. 266).

Diese intraindividuelle Bezugsnormorientierung bzw. die persönliche Ebene der Feedbackinhalte wird in der Feedbackforschung im Allgemeinen kritisch gesehen. Hattie und Timperley (2007) resümieren, dass dieses „feedback about the self as a person" (S. 96, z.B. „Great effort." oder „Good girl.") keine/kaum aufgabenrelevante Informationen enthält und damit eher nicht in mehr Anstrengung bzw. Commitment gegenüber den Lernzielen oder in einem besserem Verständnis für die Aufgabe resultiert. Die Brücke zwischen den vermittelten Informationen und den zu bewältigenden Anforderungen ist für den Lerner also schwer ersichtlich. Deshalb wird im Allgemeinen die Orientierung an den inhaltlichen Anforderungen, die kriteriumsbezogene Bezugsnorm, für Feedbackinterventionen empfohlen.

Dass die auf der persönlichen Ebene formulierten Rückmeldungen in den Untersuchungen von Schunk und Rice (1991, 1993) dennoch erfolgreich sind, bedarf der Erklärung: die elaborierten Feedbacks stimulieren hier in erster Linie die Anwendung der Lesestrategie und werden zudem nur nach richtigen Antworten gegeben. Es handelt sich quasi um eine wiederholte Erinnerung, eine verständnisförderliche Strategie umzusetzen und dabei scheint es egal, ob das nach richtigen oder fehlerhaften Antworten eingesetzt wird; der Inhalt der Rückmeldungen bliebe davon mehr oder weniger unberührt. Fällt der Strategiebezug jedoch weg, so wie in der früheren Studie der Autoren (Schunk & Rice, 1986), erweisen sich die auf der persönlichen Ebene konzentrierten Feedbacks als wirkungslos. Es wird hier lediglich eine Anstrengungs- oder Fähigkeitszuschreibung von Erfolg (z.B. „You are good at this.") rückgemeldet und damit werden eher motivationale Aspekte betont, aber keine Information angeboten, die zur Bewältigung der Anforderung des Textverstehens (hier: Hauptaussagen extrahieren) genutzt werden können. Alles in allem ist davon auszugehen, dass die elaborierten, strategiebezogenen Feedbacks von Schunk und Rice (1991, 1993) *trotz* der deutlichen individuellen Bezugsnormorientierung und der Missachtung der fehlerhaften Antworten erfolgreich sind.

Der Bezug der Rückmeldungen auf Fehler, wie es – außer bei Schunk und Rice (1986, 1991, 1993) – in den Studien umgesetzt wurde, entspricht der vorherrschenden Auffassung von Feedback und den Empfehlungen zu seiner Gestaltung, um Lernen und Leistung zu fördern. Im Rahmen des Textverstehens sind Fehler bzw. unzureichende Antworten auf Testfragen Ausdruck eines (in Teilen oder gänzlich) fehlerhaften oder fehlenden Verständnisses eines beschriebenen Sachverhaltes. Verständnisschwierigkeiten

können verschiedene Ursachen haben (vgl. Abschnitt 2.3), sie können etwa auf eine geringe Lesekompetenz oder bei prinzipiell gut ausgebildeten Fähigkeiten auch auf eher situations- bzw. textbezogene Schwierigkeiten zurückgehen. Worin die Ursachen auch liegen, um eine Kompetenz auf- und auszubauen bzw. die mentale Repräsentation eines vorliegenden Sachverhaltes/einer Situation zu korrigieren und/oder zu fördern, muss auch an den entsprechenden Defiziten angesetzt werden. Fehler signalisieren im Gegensatz zu richtigen Antworten Lerngelegenheiten bzw. einen Lernbedarf und gelten daher im Rahmen kognitiver Theorien des Lernens als essentiell.

Über die bereits genannten Merkmale hinaus fällt eine weitere Gemeinsamkeit der meisten Feedbackstudien zum Textverstehen auf:

- Die Treatments erstrecken sich über mehrere Sitzungen.

Durch die Wiederholungen der Interventionssitzungen werden mehr Lerngelegenheiten im Zusammenhang mit Feedback geschaffen, die Lernphase wird ausgedehnt und verteilt. Eine ausreichende Dauer und die Verteilung von Lerngelegenheiten werden beide als wichtige Bedingungen für das Einüben neuer Fähigkeiten angesehen. Insofern ist es plausibel, dass Interventionen für Leser mit eher geringerer Lesekompetenz (wie z.B. in Peverly & Wood, 2001; Winne et al., 1993, vgl. auch Tabelle 4) mehrteilig angelegt sind.

An den erläuterten Studien fallen allerdings auch einige Schwächen, die ebenso über den Anwendungsbereich des Textverstehens hinaus und relativ häufig im Rahmen von Feedbackstudien vorzufinden sind. Die Kritik bezieht sich auf:

- die Wahl der Kontrast-/Kontrollbedingungen,
- die Vernachlässigung des Erfassens der Leistung unmittelbar nach der Feedbackgabe und
- den Einsatz kurzer Texte bzw. Textpassagen.

Die Auswahl der Kontrastbedingungen in den Feedbackstudien ist in mehr als einer Hinsicht zu kritisieren. Zum einen fokussieren die Untersuchungen in der Regel jeweils ein bis zwei Feedbackarten und auch über die Studien hinweg wurde das Spektrum der elaborierten Feedbackarten (vgl. Tabelle 2) bisher nicht ausgenutzt. Wie oben ausgeführt beziehen sich die Rückmeldungen primär auf den Prozess, der zur Bewältigung einer Aufgabenstellung notwendig ist, und greifen dabei häufig auf (meta-)kognitive Strategien zurück. Es finden sich dagegen keine Beispiele für Rückmeldungen, die sich auf den Fehler beziehen oder die nicht am Prozess selbst, sondern an dem Thema oder der

Aufgabenstellung bzw. den Aufgabenmerkmalen ansetzen. Darüber hinaus fehlt es, wie schon Kulhavy und Stock (1989) sowie McKendree (1990) im Allgemeinen bemängeln, auch im Bereich des Textverstehens an Untersuchungen, die systematisch eine Reihe unterschiedlicher, vor allem elaborierter Rückmeldungen untersuchen. So bleiben immer Einschränkungen in der Interpretation und Generalisierbarkeit der Befunde. Zum anderen werden die Treatmentbedingungen nicht immer mit einer Kontrollbedingung ohne Feedback kontrastiert, was wiederum die Aussagekraft der Ergebnisse einschränken kann.

Der zweite Kritikpunkt bezieht sich darauf, dass die Wirksamkeit von Feedbackinterventionen typischerweise ausschließlich über eine Transferleistung in einem Posttest oder einen Prä-Posttest-Vergleich bestimmt wird. Nicht untersucht wurde dagegen bisher die Leistung unmittelbar nach der Feedbackgabe, die ein Indikator dafür ist, inwiefern die Rückmeldungen geeignet sind, um Fehler zu korrigieren. Kulhavy und Stock (1989) betrachten diesen Aspekt der Feedbackwirksamkeit als einen wesentlichen, weil er die unmittelbarste Wirkung der Intervention reflektiert. Selbstverständlich ist die Transferleistung ein nicht minder wichtiger Indikator für die Feedbackeffektivität (vgl. Abschnitt 3.4). Aber durch die Berücksichtigung der Korrekturleistung infolge einer Feedbackgabe ließe sich die Bewertung einer Feedbackintervention komplettieren.

Schließlich fällt an den berichteten Untersuchungen im Bereich des Textverstehens noch auf, dass sie häufig nur kurze Texte bzw. einzelne Passagen/Absätze eingesetzt haben. Längere Texte eröffnen dagegen aber mehr Möglichkeiten, hierarchiehöhere Prozesse wie das Herstellen von globaler Kohärenz abzuprüfen und dafür, im Falle von Schwierigkeiten, Lerngelegenheiten durch Feedback zu schaffen.

Insgesamt bestätigen die dargestellten Befunde die Annahme, dass elaboriertes Feedback im Gegensatz zu einfachen Rückmeldungen komplexe, hierarchiehöhere Aufgabenanforderungen, zu denen das Textverstehen zu zählen ist, unterstützen und fördern kann (vgl. Bangert-Drowns et al., 1991; Butler & Winne, 1995; Mason & Bruning, 2001; Schimmel, 1988; Shute, 2008). Durch elaboriertes Feedback lässt sich das Textverständnis korrigieren und/oder ausbauen bzw. die Lesekompetenz erhöhen. Hinsichtlich der Arten wirksamer Feedbackinhalte erlauben die entsprechenden Studien einen eher nur begrenzten Einblick. Denn zum einen liegen nur wenige Studien vor, die Feedback tatsächlich auf die Verstehensprozesse beim Lesen von Texten bezogen haben. Zum anderen wird dabei mit sehr ähnlichen Inhalten des Feedbacks operiert – es wird

entweder Bezug zu Lesestrategien oder den konkreten Teilprozessen des Verstehens, die für die jeweilige Aufgabenstellung relevant sind, genommen. Über diese Inhalte hinaus ist bisher kaum etwas umgesetzt worden.

Um die geringe Anzahl und den eher engen Fokus der Feedbackstudien im Bereich des Textverstehens aufzufangen, werden nachfolgend noch einige Erkenntnisse aus einem mit der Feedbackforschung verwandten Forschungsbereich, dem Prompting, ergänzt. Die Literatur zu Prompting bietet Einblicke in weitere, potentiell effektive Gestaltungsmöglichkeiten für Feedback, und die Befunde können die zuvor berichtete Nützlichkeit elaborierter Rückmeldungen für das Textverstehen bekräftigen.

Befunde aus dem Forschungsbereich Prompting

Als Prompts werden Erinnerungsstützen, Hilfestellungen oder Hinweise beim Bearbeiten von Aufgabenstellungen bezeichnet (Bannert, 2009; Wirth, 2009). Sie können zwar als eine Form des elaborierten Feedbacks eingesetzt werden (siehe Tabelle 2), dann auch meist zusammen mit Knowledge of Result oder Knowledge of Correct Result. Doch Prompts sind nicht mit Feedback gleichzusetzen, denn sie beziehen sich nicht (notwendigerweise) auf eine Aufgabenbeantwortung, sondern sind zunächst nach Wirth (2009) ein eigenständiges instruktionales Mittel, das im Verlauf eines Lernprozesses oder bei der Bearbeitung einer Aufgabenstellung eingestreut werden kann.

Die Funktion des Prompting wird darin gesehen, vorhandene, aber nicht spontan gezeigte Wissenselemente, Strategien oder Fähigkeiten bzw. Fertigkeiten zu aktivieren (Wirth, 2009). In Abhängigkeit davon, welche Art der Informationsverarbeitung aktiviert werden soll, werden kognitive und metakognitive Prompts unterschieden. Kognitive Lernhilfen werden in Beziehung zu Klassifikationen von Lernstrategien gesetzt und, implizit oder explizit, häufig in Organisations- und Elaborationsstrategien unterteilt (z.B. Berthold, Nückles & Renkl, 2007; Glogger, Holzäpfel, Schwonke, Nückles & Renkl, 2009). Metakognitive Prompts können sich dagegen auf „die Orientierung, Planung und Zielbildung vor dem Lernen, die Überwachung und Steuerung während des Lernens und die (End-) Kontrolle gegen Ende des (…) Lernens“ (Bannert, 2003, S. 14) beziehen. Darüber hinaus lassen sich kognitive und metakognitive Prompts danach unterscheiden, ob sie als konkrete Arbeitsanweisungen (z.B. „calculate first 2+2“; Bannert, 2009, S. 139)

oder als eher unspezifische, generelle Aufforderungen/Fragen gestaltet sind (z.B. „what is your plan?“; Bannert, 2009, S. 139).
Die Untersuchungssituationen sind meist computergestützte Lernumgebungen zu unterschiedlichen Domänen, typischerweise aber mathematisch-naturwissenschaftliche Themenbereiche. Im Speziellen ist die Forschung zum Nutzen von Prompting auch bezogen auf Hypertext- bzw. Hypermedia-Umgebungen zu finden (Azevedo, 2005a, 2005b, 2007; Azevedo et al., 2004; Azevedo & Jacobson, 2008; Bannert & Mengelkamp, 2008; Moos & Azevedo, 2008; Puntambekar & Stylianou, 2005). Die Wirksamkeit von Prompts wird für gewöhnlich anhand des Wissenszuwachs und/oder der Anwendbarkeit des erworbenen Wissens auf neue Problemstellungen (naher, weiter Transfer) beurteilt.

Die Befunde zum Prompting sprechen dafür, dass Prompts im Allgemeinen eine effektive Unterstützung des (selbstregulierten) Lernens darstellen (Davis, E. A., 2003; Lin & Lehman, 1999). Insbesondere kognitive Lernhilfen führen mit großer Wahrscheinlichkeit zu einer Leistungsverbesserung. Von Vorteil sind hierbei das Stimulieren von Organisationsstrategien wie Zusammenfassen, Gemeinsamkeiten und Unterschiede zwischen Elementen Herausfinden, Hauptaussagen Extrahieren (Berthold et al., 2007; Glogger et al., 2009) oder das Suchen von Überschriften (Berthold et al., 2007) sowie Elaborationsstrategien wie das Aktivieren von Vorwissen (durch Aufschreiben), das Finden von Beispielen (Berthold et al., 2007; Glogger et al., 2009), das Suchen nach Zusammenhängen und Ursachen von Ereignissen oder das Hinweisen, welche Informationen wichtig sind und gemerkt werden sollten (Thillmann, Künsting, Wirth & Leutner, 2009). Auch effektive Instruktionsprogramme oder Trainings zur Förderung des Textverstehens/der Lesekompetenz greifen im Rahmen ihrer vielfältigen Unterstützungsmaßnahmen unter anderem auf kognitive Prompts zurück, die den Lerner beispielsweise auffordern, Zusammenhänge zwischen zwei Ereignissen herzustellen, Ursache-Wirkungs-Zusammenhänge zu berücksichtigen und dergleichen (z.B. Graesser, McNamara & VanLehn, 2005 für „AutoTutor“; King, 2007).
Im Vergleich zu den kognitiven Lernhilfen führen metakognitive Prompts weniger zuverlässig zu Leistungsverbesserungen. Das heißt, trotz entsprechender Hinweise werden metakognitive Strategien eher nicht öfter, beständiger oder zielführender eingesetzt (Berthold et al., 2007; Glogger et al., 2009; Wirth, 2009). Metakognitive Prompts sprechen häufig Überwachungs- oder Regulationsprozesse an, indem sie den Lerner anregen darüber nachzudenken, welche Aspekte eines Themas oder welche

Informationen eines Textes noch nicht verstanden wurden und/oder wie Verständnisschwierigkeiten überwunden werden können.

Lin und Lehmann (1999) und van den Boom und Kollegen (2007) konnten zeigen, dass das Anregen zum Reflektieren über das Gelesene bzw. Gelernte lernförderlich sein kann. Dabei können sich die Reflexionen auf den Prozess, Merkmale der Aufgabenstellung oder sogar das Befinden des Lerners beziehen. Wenn aber erworbenes Wissen auf unähnliche Problemstellungen übertragen werden soll (weiter Transfer), sind nur noch jene Reflexionen wirksam, die die Aufmerksamkeit des Lerners auf seine Strategien oder seine Vorgehensweise beim Arbeiten in einer Lernumgebung gelenkt haben (Lin & Lehman, 1999). Dabei ergänzt Davis (2003), dass wenig handlungsleitende Prompts zur Reflexion (z.B. „stop and think“) besser geeignet sind als Anregungen, die bezogenen auf das Lernmaterial konkrete Ansatzpunkte für Reflexionen vorgeben. Auf der anderen Seite gibt es wiederum Belege, dass das Anregen von Reflexionen, die wenig spezifisch ausgerichtet sind, keine erfolgreiche Transferleistung nach sich zieht (Bannert & Mengelkamp, 2008). Die Befundlage zu Reflexionen anregenden Prompts ist insgesamt also eher uneinheitlich.

Eine weitere, sehr verbreitete Methode ist das Self-Explaining. Hierbei werden Lerner aufgefordert, sich ihre eigenen Entscheidungen (etwa in einer Lernumgebung) bzw. Schritte beim Bearbeiten einer Aufgabenstellung (z.B. Stark, Tyroller, Krause & Mandl, 2008) oder die Gegebenheiten in einem Text zum Zweck des tieferen Verständnisses des Gelesenen selbst zu erklären (z.B. Chi et al., 1994). Dieses Vorgehen hat sich als eine (hoch) effektive Maßnahme für das Verständnis schriftlicher Information und die Aneignung von Wissen erwiesen (Ainsworth & Loizou, 2003; Chi et al., 1994; Sandoval, Trafton & Reiser, 1995; Stark et al., 2008; Tyroller, 2005; Wichmann & Leutner, 2009), vor allem auch um fehlerhafte Repräsentationen oder falsche Konzepte (Misconceptions) zu revidieren (Chi, 1996). Hier zeigt sich wiederum, dass Lerner beim Self-Explaining, zumindest in komplexe Lernumgebungen mit hohen Anforderungen an den Lerner, von spezifischen Unterstützungen beim Ausführen oder Anwenden von Erklärungen profitieren (Wichmann & Leutner, 2009).

3.5.1.2 Zur Wirksamkeit von unmittelbarem und verzögertem Feedback

Bezüglich des Zeitpunkts einer Feedbackpräsentation kann im Wesentlichen zwischen einem unmittelbaren und einem verzögerten Einsatz gewählt werden (siehe Abschnitt

3.2.2). Aus der Rahmenstellung dieser Arbeit – die Anwendbarkeit der Feedbacks für einen Dynamischen Test – resultiert, dass unmittelbare Rückmeldungen (pro Item) zu konstruieren sind. Deshalb werden die Befunde zum Zeitpunkt von Feedbackinterventionen hier aus der Perspektive des unmittelbaren Feedbackgebens bewertet.

Die im vorherigen Abschnitt erläuterten Untersuchungen zur Wirksamkeit von Feedbackinhalten auf das Textverständnis/die Lesekompetenz sind auch aus der Perspektive des Zeitpunktes der Feedbackgabe beschreibbar. In den meisten dieser Studien sind die Feedbackinterventionen als unmittelbare Rückmeldungen auf die Antworten der Probanden zu den Testfragen gegeben worden (vgl. Tabelle 4). Allein in der Studie von Morrison und Kollegen (1995) wurde explizit eine Feedbackart, Knowledge of Correct Result, in einer unmittelbaren (nach jeder Antwort) und einer verzögerten Variante (am Ende der Unit) kontrastiert. Keine der beiden Varianten führte zu einer Leistungssteigerung in den Aufgaben, die das Verständnis der gelesenen Informationen abprüften. Ein Zusammenhang zwischen dem Zeitpunkt der Feedbackintervention und deren Wirksamkeit für das Textverständnis bzw. die Lesekompetenz ist insgesamt nicht offensichtlich; außer bei Morrison und Kollegen (1995) wird der zeitlichen Dimension in den Diskussionen der Originalarbeiten auch keine Bedeutung beigemessen.

Auch in der allgemeinen Einschätzung wird inzwischen wieder davon ausgegangen, dass unmittelbar gegebenes Feedback einer verzögerten Gabe im Normalfall vorzuziehen ist (Bangert-Drowns et al., 1991; Kulik & Kulik, 1988). Eine Reihe älterer, zum Teil auch jüngerer Untersuchungen hatte dementgegen einen größeren Vorteil in der Verzögerung der Feedbackgabe gezeigt (z.B. Kulhavy & Anderson, 1972; Peeck, 1979; Smith & Kimball, 2010; Webb, Stock & McCarthy, 1994). Dieses als Delay-Retention Effect bezeichnete Phänomen (vgl. Kulik & Kulik, 1988) sorgte zeitweise für große Aufmerksamkeit, weil es trotz der Verletzung eines zentralen Wirkprinzips des damals vorherrschenden behavioristischen Paradigmas der Feedbackforschung, der Unmittelbarkeit der Feedbackgabe, wirkte und dabei einen größeren Effekt auf die Leistung hatte. Als Erklärung dieses Effekts schlugen Kulhavy und Anderson (1972) die *Interference-Perseveration Hypothesis* vor, die besagt, dass durch die zeitlich verzögerte Intervention die neue (richtige) Information des Feedbacks nicht mit der Gedächtnisspur des Fehlers interferiert, weil dieser dann schon vergessen ist, und somit besser gelernt

werden kann. Diese Hypothese wurde jedoch inzwischen in zentralen Punkten widerlegt (siehe Mory, 2004, S. 755-756).

Einige Studien suggerieren, dass Lerner für Transferaufgaben eher von verzögertem Feedback profitieren, für den kurzfristigen Erfolg in den vorliegenden Aufgaben eines Treatments aber unmittelbares Feedback besser geeignet ist. Andererseits wird aber auch angenommen, dass unmittelbar gegebenes Feedback vor allem auch bei der Bearbeitung schwieriger Aufgaben von Vorteil sei (vgl. Shute, 2008).

Kulik und Kulik (1988) haben in ihrer Metaanalyse zum *Timing* von Feedback jedoch herausgearbeitet, dass verzögertes Feedback eher nur unter stark kontrollierten und künstlichen Untersuchungsbedingungen funktioniert und ansonsten eher lernhinderlich ist. Wenn verzögertes Feedback wirksam ist, dann vor allem, wenn mit ihm auch die Aufgabenstellung wiederholt dargeboten wird (Kulik & Kulik, 1988), wodurch eine erneute Lernphase vorliegt (Musch, 1999). Damit lässt sich die Bedeutung der zeitlichen Dimension relativieren und der Zeitpunkt der Feedbackgabe kann eher von der Art der Aufgabe abhängig gemacht werden (vgl. Mathan & Koedinger, 2002). Für den Kontext des Lehrens und Lernens wird generell das unmittelbare Geben von Rückmeldungen empfohlen (Bangert-Drowns et al., 1991; Corbett & Anderson, 2001). Durch das Aufschieben des Feedbacks werden die Informationen zurückgehalten, die für den Lerner bzw. den Lernprozess relevant sind (Dempsey, Driscoll & Swindell, 1993).

Welche zeitliche Variante von Lernern präferiert wird, kann nicht eindeutig beantwortet werden. Einerseits wird das unmittelbare Feedback im Vergleich zum verzögerten bevorzugt (Robin, 1978). Bei Buzhardt und Semb (2002) allerdings nur dann, wenn die Möglichkeit besteht, Testfragen zu überspringen. Andererseits wird berichtet, dass unmittelbares Feedback pro Item/Antwort nicht gemocht wird (Gaynor, 1981) oder, wenn es sich um einen wichtigen Test handelt, dem eher auch mit Besorgnis entgegengesehen wird (vgl. Buzhardt & Semb, 2002). Wiederum andere Studien zeigen, dass Lerner keine besondere Präferenz hinsichtlich des Zeitpunktes der Rückmeldung haben (Corbett & Anderson, 2001).

Die für diese Arbeit geplante unmittelbare Feedbackgabe ist vor dem Hintergrund der Forschungsliteratur also dahingehend einzuschätzen, dass sie den Empfehlungen entspricht. Trotz der Wirksamkeit des unmittelbaren Rückmeldens sind unerwünschte „Nebenwirkungen“ nicht ausgeschlossen. Shute (2008) weist darauf hin, dass die unmittelbare und damit zuverlässige Feedbackgabe dazu führen kann, dass sich der

Lerner auf die Hilfestellungen verlässt, deshalb weniger sorgfältig bzw. selbstständig arbeitet und letztlich bei einer späteren Überprüfung der Leistung, beispielsweise in einem Transfertest, wiederum das vermittelte Wissen auch nicht anwenden kann. Dieser Aspekt wird für die Diskussion der Befunde dieser Arbeit zu berücksichtigen sein.

3.5.1.3 Zur Wirksamkeit von Präsentationsmodi des Feedbacks

Die in Abschnitt 3.5.1.1 aufgeführten Untersuchungen zur Feedbackwirksamkeit im Bereich des Textverstehens griffen etwa zu gleichen Teilen auf eine personen- oder eine umgebungsvermittelte Feedbackgabe zurück (vgl. zusammenfassend Tabelle 4). Die Feedbackquelle ist wiederum konfundiert mit der Darstellungsform: werden die Rückmeldungen durch eine Person gegeben, dann sind sie fast immer in mündlicher Form, werden sie durch die Umgebung vermittelt, dann sind sie immer schriftlich dargestellt. Doch eine Kontrastierung verschiedener Modalitäten ist in keiner dieser Studien umgesetzt. Auch im weiteren Feld der Feedbackforschung ist die Fragestellung, ob oder wie stark die möglichen Feedbackeffekte vom jeweiligen Präsentationsmodus abhängen, sehr selten untersucht worden. Eine Ausnahme bildet die Studie von Kluger und Adler (1993), die unter anderem zeigt, dass die Feedbackquelle – Person oder Computer – im Allgemeinen keinen Effekt auf die Leistung oder die Motivation der Probanden hatte. Nur unter bestimmten Persönlichkeitseigenschaften der Lerner wirkt sich eine Person als Feedbackgeber negativ auf die Leistung aus (vgl. Abschnitt 3.5.2).

Daneben brachte die Studie von Kluger und Adler einen für die Autoren unterwarteten Effekt hervor: die Kontrollgruppe, die die Aufgaben ohne Feedbackintervention und unter Anwesenheit des Untersuchungsleiters bearbeitete, zeigte die höchste Leistung. Sie schnitt signifikant besser ab als die Vergleichsgruppe, die ebenfalls ohne Feedbackintervention, aber auch ohne Beisein des Untersuchungsleiters allein am Computer arbeitete. Dieses Ergebnis wird dahingehend interpretiert, dass die Anwesenheit des Untersuchungsleiters (er saß als Beobachter unmittelbar hinter dem Probanden) bei dem Lerner Verhalten im Sinne der sozialen Erwünschtheit, hier also Anstrengung oder gewissenhaftes Arbeiten, begünstigt. Unterstützung findet diese Interpretation durch die Überblicksarbeit von Guerin (1986), die belegt, dass die Anwesenheit einer anderen Person normgerechtes Verhalten begünstigt, insbesondere wenn diese Person als Beobachter auftritt und es sich dabei um den Untersuchungsleiter handelt. In der Studie von Kluger und Adler geht dieser leistungssteigernde Effekt durch

den anwesenden, beobachtenden Testleiter allerdings verloren, sobald sie/er als Feedbackgeber fungiert. In dieser Bedingung unterscheidet sich die durchschnittliche Leistung nicht von den Bedingungen, in denen ohne unter Beobachtung zu stehen allein am Computer (mit und ohne Feedback) gearbeitet wird. Das heißt, sobald der Untersuchungsleiter die Feedbackgabe übernimmt und damit den Lerner bewertet, scheinen leistungshinderliche Prozesse beim Lerner einzusetzen. Eine mögliche Erklärung liefert Comer (2007), der zeigt, dass eine Person als Feedbackquelle vor allem dann nicht gewünscht bzw. nicht akzeptiert wird, wenn es sich um negative, das heißt fehlerrückmeldende Mitteilungen handelt. Rückmeldungen, dass etwas falsch gemacht wurde, sind in der computervermittelten Variante eher akzeptiert, da hier wohl kein „Gesichtsverlust“ droht (Ashford & Cummings, 1983) und weniger negative Emotionen und/oder motivationale Schwierigkeiten zu erwarten sind.
Dieser Erklärungsansatz kann durch einen weiteren Befund der Studie von Kluger und Adler (1993) gestützt werden. Ein Teil der Probanden erhielt die Rückmeldungen nicht automatisch, sondern nur auf Nachfrage, und wiederum je nach Gruppenzugehörigkeit entweder über den Computer oder vom Untersuchungsleiter. Die Ergebnisse sprechen für die computerbasierte Vermittlung: die Lerner fordern Feedback eher dann an, wenn es über den Computer vermittelt wird. Feedback, das über einen Untersuchungsleiter einzuholen ist, wurde seltener erbeten (vgl. auch Ashford & Cummings, 1983; Karabenick & Knapp, 1988).

Die berichteten Befunde werfen zwar weitere Diskussionspunkte auf, etwa inwieweit Probanden ohne Anwesenheit eines (beobachtenden) Testleiters ihre maximale Leistungsfähigkeit überhaupt abrufen oder wie der positive Effekt der Anwesenheit eines Testleiters dennoch für Feedbackinterventionen nutzbar gemacht werden kann. Doch mit Blick auf die Gestaltung von Feedbackinterventionen sprechen die Befunde insgesamt dafür, dass sowohl auf eine Person als auch den Computer als Feedbackquelle zurückgegriffen werden kann. Wird zusätzlich eine ökonomische Untersuchungsdurchführung angestrebt, rückt die computerbasierte Variante in den Vordergrund.

Die Frage, ob die schriftliche im Gegensatz zur verbalen Darstellungsform die Feedbackwirksamkeit unterschiedlich beeinflussten, ist wie das Merkmal der Feedbackquelle eher selten untersucht worden. Einen Hinweis bietet jedoch die

Metaanalyse von Kluger und DeNisi (1996), wonach verbales Feedback in Zusammenhang mit geringeren Feedbackeffekten steht. Dabei handelt es sich vor allem um Interventionen, in denen ein Untersuchungsleiter oder eine instruierte Lehrperson die Feedbackgabe übernimmt und dieser Feedbackgeber eine prominente Stellung einnimmt. Dadurch setzen dann seitens des Lerners, so die Vermutung von Kluger und DeNisi, wieder Prozesse ein, die seine Aufmerksamkeit von der Aufgabe weg und hin zu *meta-tasks*, das heißt auf das Ich oder affektive Prozesse, geleitet. Der Grund für den leistungsminderen Effekt verbalen Feedbacks wird also in dem Hervorrufen aufgabenirrelevanten Kognitionen bzw. Mechanismen gesehen, die durch eine größere Gewichtigkeit der feedbackgebenden Person in der Intervention noch verstärkt werden. Nichtsdestotrotz sind verbale, durch eine anwesende Person gegebene Rückmeldungen im Allgemeinen nicht wirkungslos, in manchen Untersuchungskontexten kann sich diese Form der Präsentation auch als die angemessenere Variante darstellen. Eine gewisse Sorgfalt ist dann aber vor dem Hintergrund der Befunde von Kluger und DeNisi (1996) darauf zu richten, wie (weit) der Einfluss des Feedbackgebers in der Intervention gestaltet werden soll.

Dass durch eine personenvermittelte Feedbackgabe auf Seiten des Lerners leistungshinderliche Mechanismen hervorgerufen werden können, verdeutlicht sehr gut einen weiteren wichtigen Aspekt: der Feedbackprozess und schließlich die Feedbackeffektivität wird nicht allein durch die „Qualität“ der Feedbackintervention bestimmt. Feedback wirkt nicht automatisch und es ist auch kein einheitliches Phänomen (Bangert-Drowns et al., 1991; Latham & Locke, 1991). Dem Empfänger der Rückmeldungen, dem Lerner, kommt eine nicht minder bedeutsame Rolle im Prozess zu.

3.5.2 Faktor Feedbackrezeption

Lernen ist ein aktiver, konstruktiver Prozess (Kintsch, 2009) und das Nutzen von Feedback als ein Teil der Informationskette im Lernprozess setzt vom Lerner ebenso ein aktives Vorgehen voraus. „Its [feedback’s] effect on action depends on how it is appraised and what decisions are subsequently made with respect to it” (Latham & Locke, 1991, S. 224; vgl. auch Hancock, Thurman & Hubbard, 1995; Jacoby, Troutman, Mazursky & Kuss, 1984). Affektive und vor allem motivationale Bedingungen seitens

des Lerners spielen eine wichtige Rolle, die es beim Einsatz von Feedback entsprechend zu berücksichtigen gilt.

Bangert-Drowns und Kollegen (1991) haben basierend auf den Erkenntnissen ihrer Metaanalyse ein Fünfstufenmodell des Lernens formuliert (S. 217, 233f.), in dem den motivationalen Bedingungen des Lerners eine zentrale Stellung eingeräumt wird. Feedback kann erst dann wirken, wenn es bewusst und aufmerksam („*mindful*“, Salomon & Globerson, 1987) rezipiert wird. Eine grafische Darstellung des Modells findet sich bei Dempsey, Driscoll und Swindell (1993) und ist in Abbildung 2 wiedergegeben.
Das Modell spiegelt die Erkenntnis wider, dass bereits die Herangehensweise an eine Aufgabenstellung und dann auch die Rezeption und Nutzung von Feedback durch Vorwissen, Rezeptionsziele, Interessen und Selbstwirksamkeitserwartungen des Lerners beeinflusst werden. Der Informationsverarbeitungsprozess und damit die Feedbackwirksamkeit hängen also an verschiedenen Stellen von einer hinreichenden Motivation des Lerners ab (vgl. Bartholomé, Stahl, Pieschl & Bromme, 2006). Verarbeitungsprozesse können behindert oder abgebrochen werden, wenn zu hohe kognitive und/oder motivationale Anforderungen an den Lerner gestellt werden oder wenn eine achtsame Verarbeitung der Feedbackinformation durch bestimmte Umgebungsmerkmale der Intervention (z.B. presearch availability, vgl. S. 47, Fußnote 1) unterwandert wird.

Deutlicher zum Tragen kommen die motivationalen Anforderungen an Lerner in Untersuchungssituationen, die ohne entsprechende Relevanz oder Konsequenzen für die Teilnehmer sind. Einige Feedbackstudien versuchen dem zu entgegnen, indem sie beispielsweise eine (monetäre) Incentivierung für eine erfolgreiche bzw. gewissenhafte Untersuchungsteilnahme in Aussicht stellen (Morrison et al., 1995) oder ihre Studie als Pflichtbestandteil eines regulären, benoteten Unterrichts/Universitätskurses einführen. In anderen Studien wird neben dem Feedback ein weiterer Faktor eingeführt, mit dem verschiedene motivationale Bedingungen erzeugt oder kontrolliert werden können. Typische Ansätze hierfür sind das Induzieren lernrelevanter bzw. anspruchsvoller Ziele (Cervone & Wood, 1995) oder das Auswählen Lassen der zu bearbeitenden Themen/Texte (Meyer et al., 2010). Allerdings beeinflussen diese Maßnahmen die Feedbacknutzung im Allgemeinen nicht. Hier liegt die Vermutung nah, dass das Umsetzen von Feedback, zumindest sofern es sich nicht um einfache Anforderungen

handelt, ein ressourcenfordernder Prozess ist, dem nicht ohne weiteres entsprechend begegnet wird.

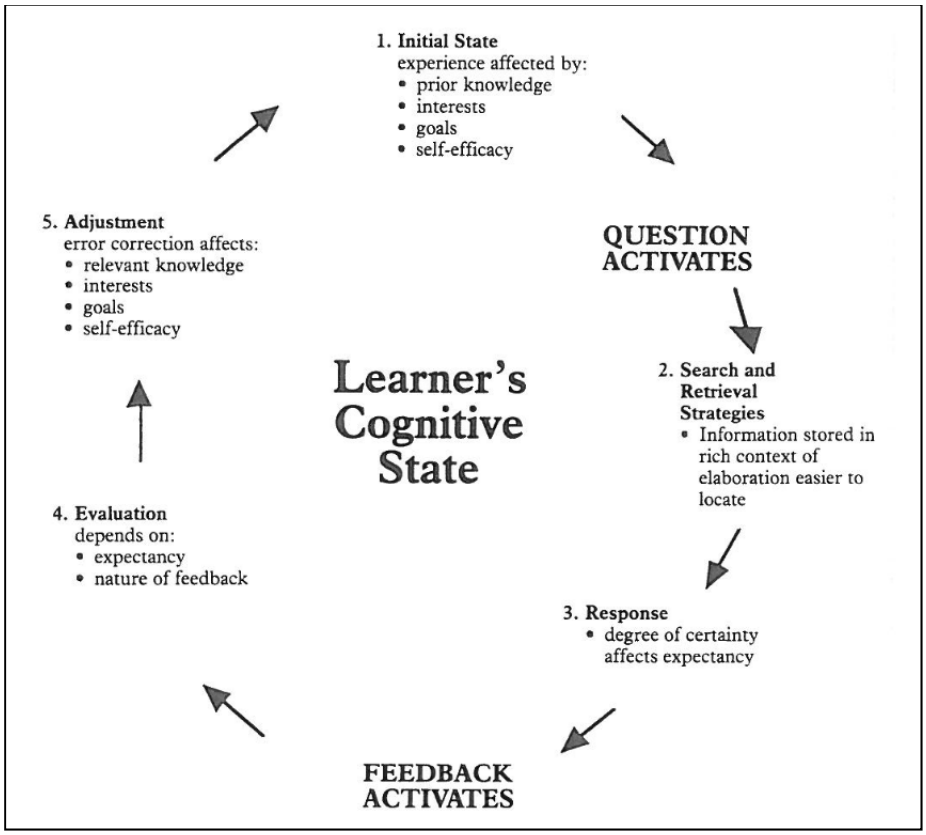

Abbildung 2 Kognitive Schritte im Lernprozess unter Feedbackgabe, Abfolge basierend auf Bangert-Drowns et al. (1991), grafische Darstellung von Dempsey, Driscoll & Swindell (1993, S. 40).

Neben den eher situativen Einflüssen auf die Testmotivation sind gelegentlich auch Personeigenschaften der Lerner hinsichtlich ihres Einflusses auf die Wirksamkeit von Feedbackinterventionen geprüft worden. Am häufigsten scheint hierbei das Vorwissen bzw. der Leistungslevel der Lerner berücksichtigt worden zu sein. Daneben sind aber auch die Konstrukte Zielorientierung, Selbstsicherheit, Selbstwirksamkeit, Feldabhängigkeit/Feldunabhängigkeit und Ängstlichkeit bzw. Angst untersucht worden.
Bezüglich des Vorwissens oder dem Leistungslevel der Lerner sprechen die Befunde dafür, dass Leistungsstärkere Feedback erfolgreicher umsetzen können (Jacoby et al., 1984). Zudem kann für sie einfaches Feedback bereits hinreichend sein (Clariana, 1990; Hanna, 1976), wohingegen leistungsschwächere Lerner mehr von elaboriertem Feedback (Hanna, 1976) oder im Rahmen der einfachen Feedbackarten eher noch von Knowledge

of Correct Result anstatt Knowledge of Result (Clariana, 1990) profitieren. Darüber hinaus scheint höheres Vorwissen eine verzögerte Feedbackgabe zu ermöglichen (Gaynor, 1981).

Die Befundlage bezüglich der Zielorientierung (Beckmann, N. et al., 2009; Davis, W., Carson, Ammeter & Treadway, 2005), Selbstsicherheit (Beckmann, N. et al., 2009; Kluger & Adler, 1993; Stake, 1982), Selbstwirksamkeit (Steele Johnson, Perlow & Pieper, 1993), Feldabhängigkeit/Feldunabhängigkeit (Roberts & Park, 1984; Whyte, Karolick, Nielsen, Elder & Hawley, 1995) und Ängstlichkeit bzw. Angst des Lerners (Frey, Stahlberg & Fries, 1986; Hansen, 1974) erweist sich dagegen uneinheitlich und wenig aussagekräftig. Dass diese motivationalen und affektiven Personeigenschaften das Lernverhalten beeinflussen können (Bartholomé et al., 2006; Efklides, 2011; Snow, 1997), ist unbestritten. Doch über die Zusammenhänge mit der Feedbackwirksamkeit können derzeit keine sicheren Aussagen getroffen werden. Gleiches gilt umso mehr für die Mechanismen des Zusammenspiels der verschiedenen Faktoren.

3.5.3 Zusammenfassung und Schlussfolgerungen für die vorliegende Arbeit

Zusammenfassend ist festzuhalten, dass sich die Effekte von Feedback auf die Leistung im Allgemeinen in einer weiten Spannbreite bewegen. Einerseits können durch Feedbackinterventionen deutliche Leistungssteigerungen erreicht werden. Andererseits können sie aber auch wirkungslos bleiben oder sogar zu Verschlechterungen der Leistungen führen. Dabei hängt die Wirksamkeit nicht nur von der adäquaten Gestaltung der Intervention selbst ab. Feedback ist Information, die einem Lerner hinsichtlich bestimmter Aspekte seiner Leistung gegeben wird, und als solche kann sie keine automatischen Reaktionen beim Lerner bewirken. Von mindestens ebenso großer Bedeutung für die Effektivität einer Feedbackintervention ist der Lerner, das heißt seine kognitiven Ressourcen und vor allem seine Motivation bzw. Bereitschaft zur Rezeption und Umsetzung der Rückmeldungen.

Die Anforderungen an die Motivation bzw. Anstrengungsbereitschaft sind bei kognitiv anspruchsvollen, komplexen Anforderungen, wie sie sich auch beim verstehenden Lesen auf Textebene ergeben, besonders zu unterstreichen. Ansätze zur Schaffung oder Erhöhung der Anstrengungsmotivation über die Induzierung entsprechender Verarbeitungsziele oder Incentivierungen haben sich in diesem Zusammenhang nicht als gewinnbringend erwiesen. Auch Untersuchungen zu möglichen differenziellen Effekten

bestimmter kognitiver und motivational-emotionaler Personmerkmale (z.B. Vorwissen, Selbstwirksamkeitserwartungen, Testangst) auf die Feedbacknutzung lassen bisher keine eindeutigen Schlussfolgerungen zu. Nichtsdestotrotz empfiehlt es sich, diese möglichen Einflüsse auf die Feedbackwirkung in Untersuchungen zu berücksichtigen bzw. zu kontrollieren.

Die meisten Erkenntnisse der Feedbackforschung liegen bezüglich des Faktors der Feedbackgestaltung (Feedbackinhalt, Zeitpunkt, Präsentationsmodus) vor. Dabei kommt dem Feedbackinhalt, neben der Kontrolle für eine mögliche Vorabeinsicht in Rückmeldungen (vgl. Abschnitt 3.5.1), der entscheidende Einfluss auf die Effektivität der Interventionen zu. Welche Art des Feedbackinhalts potentiell nützlich und/oder anderen Inhalten vorzuziehen ist, hängt auch von der Art der Anforderung der entsprechenden Aufgabenstellung ab. Für kognitiv anspruchsvolle und komplexe Aufgabenstellungen wird das Potential von Feedback in den elaborierten, nicht den einfachen Feedbackarten vermutet. Allerdings beruhen die bestehenden Metaanalysen und Überblicksarbeiten zu Effekten von Feedback entweder auf Studien mit vergleichsweise weniger voraussetzungsvollen Aufgaben (v.a. Wissen- und Gedächtnisaufgaben) oder sie haben keine Differenzierung hinsichtlich der Art der Aufgabenanforderungen vorgenommen (vgl. Abschnitt 3.5.1).
Darüber hinaus liegen auch nur wenige Studien vor, die Feedback für verstehendes Lesen auf Textebene untersucht haben. Deren Befunde lassen den Schluss zu, dass einfaches Feedback (Knowledge of Result, Knowledge of Correct Result und Answer-until-Correct) im Allgemeinen keinen Nutzen auf das Textverständnis/die Lesekompetenz ausübt. Elaboriertes Feedback, das in der Regel Knowledge of Result inkludiert, hat sich dagegen als potentiell wirksame Intervention für diesen Fähigkeitsbereich erwiesen. Verständnisschwierigkeiten auf der Textebene bedürfen im Allgemeinen also eher weiterführende Informationen, ausführlichere Hilfestellungen. Diese Interpretation ist auch mit den Arbeiten in Einklang zu bringen, die zeigen, dass Verständnisschwierigkeiten auf Seiten des Lesers sehr häufig mit Schwierigkeiten im Herstellen von Inferenzen verbunden sind (vgl. Abschnitt 2.3). Dass defizitäre Fähigkeiten der Textverarbeitung durch das Rückmelden der Inkorrektheit einer Antwort auf eine Testfrage und/oder das Nennen der richtigen Antwort nicht hinreichend unterstützt werden, sondern umfangreichere Hilfestellungen benötigen, erscheint schlüssig.

In den Studien, die elaboriertes Feedback für Verstehensanforderungen beim Lesen von Texten gegeben haben, sind zwei inhaltliche Schwerpunkte effektiver Interventionen auszumachen. Die Feedbackinterventionen haben meistens kognitive und metakognitive Lesestrategien vermittelt – entweder in Abstimmung mit einem zusätzlich vermittelten Lesestrategietraining oder als strategieanregende Rückmeldungen ohne zusätzliches Training. Außerdem sind Rückmeldungen eingesetzt worden, die konkret an den für die Bewältigung einer Aufgabenstellung notwendigen Teilprozessen des Textverstehens, konkret dem Ziehen von Inferenzen, ansetzen. Das Feedback erklärt hier, warum bestimmte Textinformationen relevant sind und wie die Inferenzen zu ziehen sind.
Dabei zeichnen sich die Interventionen dadurch aus, dass ihre elaborierten Feedbackinhalte, in Übereinstimmung mit den allgemeinen Empfehlungen (vgl. Abschnitt 3.2.1), auf einer kriteriumsorientierten Bezugsnorm basieren. Zusätzliche Rückmeldungen auf der individuellen Bezugsnorm, etwa Lob oder Aufmunterung, sind dabei nicht ausgeschlossen, insbesondere in den Einzel- oder Kleingruppensettings mit einer Person als Feedbackgeber. Aber die eigentlich interessierenden, elaborierten Feedbacks orientieren sich an dem jeweiligen Leistungskriterium und knüpfen darüber hinaus in erster Linie an Fehlern an.

Die Befunde zur Wirksamkeit von Prompts, die typischerweise zur Förderung des (selbstregulierten) Lernens eingesetzt werden, unterstützen die Erkenntnisse zum Nutzen elaborierter Feedbacks. Effektive Prompts beziehen sich in der Regel auf kognitive und metakognitive Lernstrategien (z.B. zusammenfassen, Gemeinsamkeiten finden, Beispiele finden, Vorwissen aktivieren) und unterfüttern damit die Feedbackstudien mit (meta-) kognitiven Lesestrategien. Zudem zeigen die Studien zum Prompting, dass kognitive Hilfen im Allgemeinen besser geeignet sind, das Lernen zu unterstützen, als metakognitive Prompts.
Auch in der Feedbackliteratur wird davon ausgegangen, dass spezifischeres Feedback effektiver für die Korrektur eines Fehlers ist als weniger spezifische Rückmeldungen (vgl. Abschnitt 3.2.1). Es kann also vermutet werden, dass elaboriertes Feedback in Form kognitiver Hinweise (z.B. Lesestrategien) im Allgemeinen mit größerer Wahrscheinlichkeit zu einer Verbesserung der Leistung führt als metakognitive Hinweise.

Viele der (effektiven) Feedbackinterventionen im Bereich des Textverstehens basieren zudem auf mehrmaligen Treatmentsitzungen, in Einzel- oder Kleingruppensettings, in

denen typischerweise auch eine Person (z.B. ein Untersuchungsleiter oder eine instruierte Lehrperson) die Feedbackgabe durchführt. Eine experimentelle Variation von Präsentationsmodi und/oder Dauer der Intervention wurde in keiner dieser Studien durchgeführt und ist auch über den Bereich des Textverstehens hinaus kaum untersucht.
Es liegen jedoch Hinweise vor, dass eine personengebundene, mündliche Feedbackgabe auch mit kritischen Aspekten behaftet sein kann, insbesondere wenn sich das Feedback auf Fehler des Lerners bezieht. Das Problem besteht im Wesentlichen darin, dass sich dabei für den Lerner eine Bewertungssituation ergibt, die aufgabenirrelevante Kognitionen und/oder negative Emotionen hervorrufen und zu motivationalen Schwierigkeiten führen kann. Mit der alternativen, computervermittelten Feedbackpräsentation ist diese Problematik zu umgehen. Allerdings hat diese Form wahrscheinlich auch zur Folge, dass die Testanden offenbar unter ihrer maximalen Leistungsfähigkeit zurückbleiben. Die Anwesenheit eines (beobachtenden) Testleiters wirkt sich förderlich auf die Testleistung aus, aber sobald der Testleiter auch zum Feedbackgeber wird, geht dieser Vorteil verloren. Letztlich sprechen die Befunde dafür, dass sowohl auf eine Person als auch den Computer als Feedbackquelle zurückgegriffen werden kann, und damit der Untersuchungskontext bei der Wahl des Präsentationsmodus von Feedback in den Vordergrund rücken kann.

Die referierten Studien zur Feedbackwirksamkeit auf das Textverstehen sind in manchen Aspekten auch zu kritisieren (z.B. Einsatz eher kürzerer Texte/einzelner Passagen). Dabei sind insbesondere zwei Aspekte herauszustellen, die sich als fruchtbar für die Feedbackforschung erweisen könnten: Zum einen wurde zur Beurteilung der Wirksamkeit der Feedbackinterventionen bisher nicht die Leistung unmittelbar nach der Feedbackgabe untersucht. Diesem Indikator wird in der Literatur viel Bedeutung beigemessen, da er erfasst, inwieweit eingesetzte Rückmeldungen von den Lernern zur unmittelbaren Fehlerkorrektur genutzt werden können. Die Erfassung dieses Indikators könnte die Beurteilung der Wirksamkeit von Feedbackinterventionen, die normalerweise ausschließlich anhand von Transferleistungen gemessen wird, facettenreicher ermöglichen.
Zum anderen fällt auf, dass die Untersuchungen in der Regel jeweils einen engen Fokus auf ein bis zwei (elaborierte) Feedbackarten gelegt haben. Auch über die Studien hinweg zeigt sich, dass aus dem eigentlich breiten Spektrum an diskutierten elaborierten Feedbackarten bisher nur ein kleiner Ausschnitt für die hierarchiehöheren Anforderungen

der Lesekompetenz umgesetzt wurde. Im Wesentlichen beschränken sie sich auf Hinweise, die auf die Lösung bzw. die richtige Ausführung der zur Bewältigung einer Aufgabe erforderlichen Teilschritte ausgerichtet sind.

Das Ausloten der Effektivität mehrerer, inhaltlich verschiedener Feedbacks würde die Erkenntnisse und Möglichkeiten der Feedbackinterventionen in dem Bereich des Textverstehens/der Lesekompetenz auf eine breitere Basis stellen und weitere Perspektiven für die Unterstützung der Prozesse der Bedeutungskonstruktion beim Lesen schaffen (z.B. gestufte Hilfesysteme, differenzielle Unterschiede für gute/schwache Leser). Aber welche der in der Literatur auffindbaren elaborierten Feedbackarten empfehlen sich für das Textverstehen? Welche anderen Inhalte als der Bezug zu einer (vorher vermittelten) Lesestrategie sind hierfür sinnvoll und welches Ausmaß der Effektivität auf das Textverständnis/die Lesekompetenz ließe sich jeweils vermuten?

Im Prinzip lassen sich alle elaborierten Arten (vgl. Tabelle 2) auch auf das Textverstehen übertragen: das Aufzeigen des Fehlers, Erklären, warum eine Antwort falsch ist (z.B. Chronologie von Ereignissen vertauscht, gegenläufige Informationen im Text), Hinweise zur Anregung notwendiger kognitiver und/oder metakognitiver Prozesse (z.B. das Herstellen einer geforderten Inferenz, Integrieren konfligierender Informationen, Überwachung des Verständnisses), Hinweise zu Aufgabenmerkmalen (z.B. Aufgabenstamm, relevante Textstellen, Art der Anforderung) und die Vermittlung relevanten Hintergrundwissens (v.a. bei Sachthemen), im Speziellen auch die Vermittlung von Wissen zur Korrektur falscher Konzepte (Misconceptions).
Jede Art von Feedback kann auf die Anforderungen des Textverstehens angepasst werden. Die Ansatzpunkte sind verschiedene, doch jede Variante vermag einen Beitrag zur Stärkung des Verständnisses eines Sachverhaltes (Situationsmodells) zu leisten. Aber in Abhängigkeit vom Kontext bieten sich hinsichtlich der Umsetzbarkeit bestimmte Optionen eher an als andere. Beispielsweise erscheint das Vermitteln zusätzlichen, relevanten Wissens im Rahmen einer Lernumgebung zum Zweck des Faktenlernens eher angezeigt als etwa beim Lesen einfacher Erzähltexte. Für eine Anforderung zum problembasierten Lernen mit offenen Aufgabenstellungen könnte dagegen unter anderem das Hinweisen auf schwierigkeitsgenerierende Aufgabenmerkmale von größerer Relevanz sein.

Der Kontext der vorliegenden Arbeit wird maßgeblich durch ihren Zweck bestimmt, wirksame Feedbackinterventionen für einen geplanten Dynamischen Test der Lesekompetenz zu identifizieren. Deshalb werden zunächst einige Vorbemerkungen zum Kern Dynamischer Tests und den daraus resultierenden Rahmenbedingungen für zu konstruierende Feedbackinterventionen gemacht. Danach werden Schlussfolgerungen zu den für die vorliegende Arbeit relevanten Feedbackarten angestellt.

Vorbemerkungen zum Untersuchungskontext dieser Arbeit

Das Rahmenkonzept der Arbeit ergibt sich aus der geplanten Entwicklung eines dynamischen Lerntests der Lesekompetenz (vgl. Dörfler et al., 2009; Dörfler, Golke & Artelt, 2010). Dynamische Tests bauen auf der Grundidee der Zonen der aktuellen und der proximalen Entwicklung nach Vygotsky (1964) auf. Demnach sollten für die Bewertung der Leistungsfähigkeit eines Individuums sowohl das aktuelle Kompetenzniveau als auch die Entwicklungspotenz, die „Zone der proximalen (nächsten) Entwicklung“, erfasst werden (vgl. Sternberg & Grigorenko, 2002). Dynamische Tests greifen diesen Gedanken auf, indem sie zusätzlich zur Kompetenztestung spezifische Interventionen (Anregungen, Hilfestellungen) zur Förderung einer Kompetenz bzw. einer Teilfähigkeit implementieren. Das Ausmaß, in dem ein Lerner auf die angebotenen Interventionen hin seine Leistung verbessern kann, wird als individuelles Lernpotential beschrieben (Beckmann, J. F., 2001; Lussier & Swanson, 2005).

Dynamische Tests liegen in zwei Varianten vor (Dillon, 1997): a) ein „Langzeitlerntest“ (oder: *test-train-test design*), in dem ein Trainingsblock zwischen zwei, normalerweise an verschiedenen Tagen durchgeführten Testungen (Prä- und Posttest) eingebettet ist, und b) ein „Kurzzeitlerntest“ (*train-within-test design*), in dem im Verlauf der Kompetenztestung Hilfestellungen unmittelbar auf (falsche) Antworten gegeben werden. Testung und Intervention sind hier also eng miteinander verzahnt und wechseln sich ab. Die Hilfestellungen sind typischerweise Feedbacks und der Kurzzeitlerntest wird in der Regel in einer Sitzung durchgeführt (Guthke & Wiedl, 1996; Sternberg & Grigorenko, 2002).

Der über diese Arbeit hinausgehende, noch zu entwickelnde Dynamische Test der Lesekompetenz ist als Kurzzeitlerntest geplant. Um die Testprozedur inklusive Feedbackgaben zuverlässig umsetzen zu können, wird der Test computerbasiert

durchgeführt werden. Zudem wird dabei auf geschlossene Antwortformate zurückgegriffen, um das Feedback ad hoc auf die Antworten geben zu können, bei offenen Antwortformaten wäre das nur beschränkt und sehr viel aufwändiger möglich (vgl. z.B. Lenhard, Baier, Hoffmann & Schneider, 2007).

Schlussfolgerungen für die Feedbackinterventionen

Durch die Ausrichtung der Feedbacks auf den geplanten Dynamischen Test der Lesekompetenz in der Variante eines Kurzzeitlerntests ergeben sich spezifische Anforderungen an den Kontext des Tests und insbesondere an die zu konzeptualisierenden Feedbackinterventionen:

- Feedback wird, wie der Test, computerbasiert gegeben.
- Feedback wird unmittelbar für Antworten in den Testaufgaben präsentiert.
- Feedback wird nach Fehlern (falschen Antworten) gegeben.
- Es wird auch die Leistung nach der Feedbackgabe erfasst, indem zweite Antwortversuche in den Test implementiert werden.
- Es werden kein zusätzlicher Trainingsblock bzw. andere umfangreiche Instruktionsmaßnahmen implementierbar sein und die Rückmeldungen dürfen (deshalb) eher nicht zu lang oder komplex sein.

Zunächst ist festzuhalten, dass jede Form des elaborierten Feedbacks in jedem Fall auch Knowledge of Result enthalten sollte. Das Feedback bezieht sich auf falsche Antworten und entsprechend empfiehlt es sich, jede Rückmeldung mit der Information zu beginnen, dass ein Fehler vorliegt (z.B. „Das ist falsch. …“). Diese Kombination wird in der Literatur empfohlen (Kulhavy & Stock, 1989) und in der Regel auch so umgesetzt.
Darüber hinaus sind vor dem Hintergrund der aufgeführten Rahmenbedingungen bestimmte elaborierte Feedbackarten auszuschließen. Dazu wird die Vermittlung relevanten Hintergrundwissens gezählt. Über diese Feedbackart könnten zum einen Informationen zum Thema des Textes gegeben werden, wodurch die Rückmeldungen eher umfangreich ausfallen und vermutlich nach dem „Gießkannenprinzip“ ablaufen würden. Eine Passung der einzelnen Rückmeldung auf die entsprechende Aufgabenstellung scheint hier weniger gut möglich und darin wird das Risiko gesehen, dass die Information beliebig wirkt. Zum anderen könnte mit dieser Feedbackart jeweils auch eine Informationseinheit vermittelt werden, die genau die Inferenz betrifft, die in

einer Aufgabe erfragt wird. Allerdings grenzt diese Umsetzung an ein Vorsagen oder Suggerieren der richtigen Antwort. Wenn eine Aufgabe nach einer Inferenz fragt, fungiert die Vermittlung des Hintergrundwissens, das es bräuchte, um diese Inferenz herzustellen, beinahe wie das Erklären des Zusammenhangs. Die richtige Lösung ist dann wahrscheinlich schon impliziert, zumindest bei geschlossenen Aufgabenformaten könnte die richtige Antwort dann leicht zu erhalten sein. Zudem kann davon ausgegangen werden, das haben die Studien zu den Ursachen von Verständnisschwierigkeiten gezeigt (vgl. Abschnitt 2.3), dass fehlendes Hintergrundwissen im Allgemeinen eher nicht die Ursache für unzureichende (elaborative) Inferenzen ist (den Einsatz altersangemessener Themen/Texte vorausgesetzt).

Ein Spezialfall ist das Feedback, das Informationen (Wissen) zur Korrektur von Misconceptions bietet. Hier ist die Vermittlung von Wissen wahrscheinlich die sinnvollste Intervention. Aber diese Arbeit fokussiert nicht das Erfassen falscher Konzepte, dafür ist ein anderes Vorgehen, ein anderes Untersuchungsdesign erforderlich. Insofern ist auch diese Variante elaborierten Feedbacks für die vorliegende Arbeit ausgeschlossen.

Aufgrund der Rahmenbedingungen wird außerdem Feedback in Form von Hinweisen zu schwierigkeitsgenerierenden Aufgabenmerkmalen ausgeschlossen. Die Testaufgaben sind in ihrem Format identisch und auch hinsichtlich ihrer inhaltlichen Anforderungen sind sie eher homogen gehalten (v.a. inferenzielle Anforderungen). Die Rückmeldungen würden sich auf einige wenige Merkmale (Aufgabenstamm oder Art der Anforderung) beziehen können, die Formulierungen der Rückmeldungen wären einander womöglich sehr ähnlich. Prinzipiell ist dieses Vorgehen zur Förderung der (Test-)Leistung im Bereich des Textverstehens/der Lesekompetenz denkbar, andere Feedbackarten versprechen aber mehr Potential.

Zu der Rückmeldung über Aufgabenmerkmale wird auch das Aufzeigen der zur Beantwortung einer Aufgabenstellung relevanten Textinformationen gezählt. Diesbezüglich kann aber wiederum mit Blick auf die Erkenntnisse zu den Ursachen von Verständnisschwierigkeiten vermutet werden, dass es sich um einen weniger erfolgversprechenden Ansatz handelt. Es zeigt sich zwar, dass ein Aspekt bei Verständnisschwierigkeiten darin liegen kann, dass die Informationen, die zum Herstellen einer Inferenz erforderlich sind, im Text nicht identifiziert werden können. Aber auch ein Aufzeigen der relevanten Textstellen trägt bei schwachen Lesern eher nicht dazu bei, dass die entsprechenden Sinnzusammenhänge daraufhin herstellbar sind (vgl. Abschnitt 2.3).

Ferner ist bei dieser Art des Feedbacks, wenn es allein gegeben wird, zu hinterfragen, ob Leser daraus lernen können. Die Textstellen würden ihnen durch den Feedbackgeber bzw. den Computer markiert. Der Leser wird dabei aber kaum Prinzipien/Strategien ableiten können, die auf folgende Aufgaben übertrag sind. Diese Annahme wird auch durch die Ergebnisse von Winne und Kollegen (1993) gestützt, wonach das Markieren der relevanten Informationen keine Leistungssteigerung bezüglich des Herstellens von Inferenzen erbringen konnte (vgl. Abschnitt 3.5.1.1).

Zu den vor dem Hintergrund des Untersuchungskontextes potentiell nützlichen Feedbackarten werden jene gezählt, die deutlich auf die Konstruktionsprozesse beim Textverstehen ausgerichtet werden können. Dazu gehören das Aufzeigen des Fehlers, das Erklären des Fehlers sowie Hinweise zur Anregung notwendiger kognitiver und/oder metakognitiver Prozesse.

Das Aufzeigen oder Benennen eines Fehlers und das Erklären desselbigen sind zwar theoretisch unterscheidbare Feedbacks. Die Fehlererklärung stellt sozusagen die Weiterführung des Nennens eines Fehlers dar. Doch beide Varianten sind sich (bezogen auf das Textverstehen) einander auch sehr ähnlich. Auf der Grundlage empirischer Befunde ist eine Einschätzung, inwieweit das Nennen und/oder Erklären eines Fehlers als eigenständige Rückmeldung geeignet ist, kaum möglich. Entsprechende Studien in diesem Bereich sind nicht bekannt. Aber die Fehlererklärung wird vereinzelt als Teil effektiver, umfangreicherer Feedbackstrategien eingesetzt. Für das Textverstehen etwa in der Studie von Winne und Kollegen (1993; vgl. Tabelle 4).

Darüber hinaus erscheint das Nennen des Fehlers im Rahmen eines Kompetenztests, der wie in dieser Arbeit mit geschlossenem Antwortformat arbeitet, weniger plausibel umzusetzen. Denn bei einem geschlossenen Antwortformat zeigt die ausgewählte, aber falsche Antwortalternative (Distraktor) im Prinzip schon den Fehler bzw. ein entsprechendes Feedback der Fehlernennung würde kaum andere Informationen beinhalten als die Bestätigung, dass zum Beispiel die Inferenz, die der Distraktor suggeriert, unzulässig ist.

Anders stellt es sich für die Fehlererklärung dar. Dem Leser zu erläutern, warum die gewählte Antwort, die im Normalfall Ausdruck seines (allerdings fehlerhaften) Verständnisses ist, falsch ist, verspricht für den Leser einen (Erkenntnis-)Gewinn. Es erscheint auch intuitiv, diese Information (als erstes) zu vermitteln. Aus der Perspektive der Theorien des Textverstehens besteht natürlich eine „Lücke“ zwischen dem

Erkennen/Verstehen, warum das eigene mentale Modell eines Textes/Sachverhalts in einem bestimmten Aspekt falsch ist, und der Neukonstruktion bzw. dem Korrigieren des angesprochenen fehlerhaften Teils des Modells. Dieses Feedback wirkt dann, wenn es den Leser anregt, durch Überlegen und unter Zuhilfenahme des Situationsmodells und/oder dem erneuten Nachlesen im Text (oder anderen Strategien) das Modell zu korrigieren und damit die richtige Antwort auf die Frage zu erlangen. Dabei handelt es sich um einen voraussetzungsvollen Prozess, zumindest in den Fällen, in denen der Leser nicht schon vor der ursprünglichen Antwort zwischen alternativen Antworten geschwankt hat. Für eine Transferwirkung des Feedbacks ist es notwendig, dass der Leser aus den Fehlererklärungen für die einzelnen falschen Antworten ein Prinzip bzw. bestenfalls die Prinzipien (nicht jedem Distraktor im Test wird ein und dieselbe Art des Fehlers zugrunde liegen) erkennt, entsprechende Strategien zur Bewältigung der fehlerhaften Prozesse verfügbar hat und diese auf die neuen Aufgabenstellungen überträgt.

Als Weiterführung einer Fehlererklärung kann das Feedback gesehen werden, das Hinweise gibt, die dem Lerner spezifisch die kognitiven Schritte anbieten, die zur Bewältigung der gestellten Anforderung notwendig sind. Bezogen auf Verständnisfragen bedeutet dies, dass das Feedback die durch die konkrete Aufgabe erfragte, aber zunächst nicht herstellbare Inferenz stimuliert. Es gibt Hinweise, wie die Sinnzusammenhänge zu konstruieren sind. Das kann das Verknüpfen von Ereignissen, die Ordnung der zeitlichen Abfolge von Ereignissen, das Überprüfen der Ursache einer Situation oder der Wirkung eines Ereignisses betreffen. Um die Hinweise erfolgreich nutzen zu können, muss der Leser die kognitiven Schritte umsetzen. Die Rückmeldungen zeichnen den Weg vor, aber die Umsetzung obliegt dem Lerner.

Die Nützlichkeit dieser Art der Unterstützung lässt sich am klarsten aus der Literatur ableiten: nicht zu wissen, wann und wie die Textinformation mit dem Vorwissen zu verknüpfen ist, ist ein wesentlicher Grund für Verständnisschwierigkeiten, die auf defizitäre Fähigkeiten von Lesern in der Textverarbeitung zurückgehen (vgl. Abschnitt 2.3). Da liegt es auf der Hand, genau an dieser Schwachstelle anzusetzen und zu vermitteln, wie die geforderte Inferenz zu ziehen ist.

Im Bereich des Textverstehens ist dieses Feedback in vergleichbarer inhaltlicher Ausrichtung schon von Winne und Kollegen (1993) erfolgreich eingesetzt worden. Deren Vorgehen hebt sich aber dadurch ab, dass die Rückmeldung zusätzliche Bestandteile enthielt (Markieren und Aufzeigen relevanter Textinformationen, Zeigen und Erklären,

wie Inferenz zu ziehen ist, vgl. Abschnitt 3.5.1.1). Im Gegensatz dazu ist das Feedback hier aufgrund der Rahmenbedingungen der Untersuchung kürzer, prägnanter zu gestalten, in ein bis zwei Sätzen, fokussiert auf das Herstellen der Inferenz. Dafür ist die Essenz der Fragestellung bzw. der geforderten Inferenz zu bestimmen und in gut verständlicher Form als „kognitive Anleitung“ zu formulieren.

Feedback zur Unterstützung des Textverstehens kann sich auch auf die metakognitive Ebene beziehen. Entsprechende Hilfestellungen im Bereich des Promptings von selbstreguliertem Lernen beziehen sich auf die verschiedenen Phasen der Selbstregulation, also die Planung, Überwachung und Regulation des Lernprozesses. Die Übertragung für Feedbacks im Bereich des Textverstehens liegt nah und metakognitive Feedbacks wurden auch schon dafür eingesetzt (vgl. Abschnitt 3.5.1.1). Diese beziehen sich eher auf die Planung oder Regulation der Aktivitäten, meist bezogen auf vorher vermittelte Lesestrategien. Der Einsatz metakognitiver Prompts ohne zusätzliche Strategievermittlung setzt dagegen voraus, dass entsprechende Fähigkeiten seitens des Lesers vorhanden sind, die durch die Rückmeldung angeregt und umgesetzt werden können.

Für erfolgreiches Textverstehen wird insbesondere die Komponente der Verstehensüberwachung unterstrichen (vgl. Abschnitt 2.3). Ein Feedback, das daran anknüpft, regt den Leser dazu an, sein Verstehen in Bezug auf die Fragestellung beim (Nach-)Lesen im Text zu überwachen und sich mit dem Text, der Aufgaben bzw. seinem konstruiertem Situationsmodell auseinanderzusetzen. Die Effektivität der Lernhilfe ist daran gekoppelt, dass Leser angeregt durch den Hinweis ihr eigenes Verständnis bezüglich der vorliegenden Fragestellung prüfen und unter Einsatz der ihnen vorhandenen (meta-)kognitiven Strategien inkohärente Anteile ihrer Bedeutungsrepräsentation des Textes erkennen und korrigieren können.

Auch ein Ansetzen an den Komponenten der Planung und Regulation ist im Rahmen eines Lesekompetenztests mit Feedback prinzipiell umsetzbar. Doch im Rahmen des vorliegenden Untersuchungskontextes erscheint ein Bezug des Feedbacks zu Planungsaktivitäten des Leseprozesses (wie i.d.R. Setzen von Zielen, Formulieren von Fragen an Text) weniger notwendig und dabei für den Lerner schlüssig als Feedback vermittelbar. Hinsichtlich der Regulationskomponente kann davon ausgegangen werden, dass sich die entsprechenden Aktivitäten an eine erfolgreich angestoßene Überwachung anschließen, aber natürlich auch nur in dem Maß, in dem der Leser über entsprechende

Strategien und das Wissen, wann diese einzusetzen sind, verfügt. Eine Vermittlung entsprechender Regulationsstrategien über Feedback, aber ohne zusätzliche Instruktionen der Strategievermittlung erscheint weniger erfolgversprechend. Daher wird die Praktikabilität metakognitiver Hinweise als Feedbackintervention in erster Linie in der Komponente der Verstehensüberwachung gesehen.
Metakognitive Prompts haben sich im Vergleich zu kognitiven Hinweisen im Allgemeinen als weniger effektiv erwiesen (vgl. Abschnitt 3.5.1.1). Die Voraussetzungen an das Können bzw. das Strategiewissen des Lesers und die Anforderungen, die ein metakognitiver Hinweis als Feedback nach einer falschen Antwort an den Leser stellt, um daraus profitieren zu können, sind vergleichsweise höher als bei kognitiven Hinweise zum Ziehen einer Inferenz (vgl. Butler & Winne, 1995). Zwischen dem metakognitiven Prompt und dem Feedback Fehlererklärung fällt ein Vergleich der Nützlichkeit aus theoretischer Perspektive schwerer. Ihre jeweiligen Vorteile und Herausforderungen scheinen sich beinahe auszugleichen – das metakognitive Feedback geht nicht auf den Fehler ein, regt dafür aber das Verstehen bzw. die Überwachung dieses an. Die Fehlererklärung vermittelt dem Leser Schwachstellen seiner Situationsmodellbildungen, geht darüber hinaus aber nicht auf den Prozess der Korrektur ein oder gibt Anregungen dahingehend.

Die bisherigen Überlegungen zur Umsetzbarkeit und Nützlichkeit elaborierter Feedbackarten im Kontext der vorliegenden Arbeit haben drei elaborierte Arten herausgestellt, die unter dem speziellen Untersuchungskontext und den daraus resultierenden Rahmenbedingungen der Feedbackinterventionen als potentiell wirksame Unterstützungen des Textverstehens eingeschätzt werden: die Fehlererklärung, kognitive Hinweise zum Herstellen einer Inferenz und metakognitive Prompts zur Anregung der Verstehensüberwachung. Alle drei Feedbackarten haken an verschiedenen Stellen des Verständnisses bzw. des Verstehens ein und können ausgehend von den verschiedenen Ansatzpunkten mehr oder weniger spezifisch zur Unterstützung des Textverständnisses/der Lesekompetenz führen.

Experiment 1 (Fokus: Kontrastierung von Feedbackinhalten)

Das Ziel des Experiments ist die Untersuchung der Effektivität verschiedener Feedbackarten auf das Textverstehen/die Lesekompetenz unter den spezifischen Rahmenbedingungen, die sich aus dem Testprinzip eines Dynamischen Kurzzeitlerntests ergeben (vgl. Abschnitt 3.5.3). Aufgrund der im vorherigen Abschnitt 3.5.3 angestellten Überlegungen stehen drei elaborierte Feedbackarten zur Verfügung, für die ein Potential zur Steigerung des Textverständnisses bzw. der Lesekompetenz angenommen werden kann: Fehlererklärung, metakognitiver Prompt und Inferenzprompt. Hier sollen alle drei Arten experimentell untersucht werden, um so Erkenntnisse über die Effekte verschiedener, wahrscheinlich unterschiedlich wirkungsvoller Rückmeldungen zu erhalten. Gemessen wird die Effektivität der Feedbackinterventionen nicht nur anhand von zwei Posttests (ein unmittelbarer und ein Follow-up), sondern auch anhand der Leistung während der Feedbackintervention, die zweite Antwortversuche vorsieht (vgl. Abschnitt 3.5.3) und damit in Erst- und Zweitantworten unterteilt werden kann (vgl. Abschnitt 3.4). Kontrastiert werden die drei elaborierten Feedbackinterventionen mit zwei Kontrollbedingungen: mit einer feedbackfreien Testbedingung und mit einer Intervention anhand des einfachsten Feedbacks, Knowledge of Result. Das hat, wie in Abschnitt 3.5.1.1 ausgeführt, zwar keine leistungssteigernde Wirkung beim Textverstehen, bietet dabei aber aufgrund derselben Feedbackprozedur wie die elaborierten Feedbackinterventionen einen zusätzlich nützlichen Vergleich zur Bestimmung derer potentiellen Effekte. Die Auswirkungen der elaborierten Feedbackinterventionen auf die Leistung stehen im Vordergrund.

Da aufgrund der computerbasierten Untersuchung auch die Bearbeitungszeiten zur Verfügung stehen, werden diese als ein grober Indikator für das Arbeitsverhalten im Test ebenfalls genutzt und auf einen möglichen Feedbackeffekt hin untersucht. Mit Bezug auf die in Abschnitt 3.5.2 dargestellten möglichen (negativen) Einflüsse bestimmter Personmerkmale auf die Feedbackrezeption und unter Berücksichtigung des speziellen Kontextes dieser Untersuchung, soll zusätzlich der Frage nachgegangen werden, ob die Feedbackinterventionen ihrerseits eine Erhöhung der Testangst bewirken. Schließlich werden noch die Auswirkungen der Feedbacks auf die Einschätzungen der Probanden hinsichtlich der subjektiv wahrgenommenen Nützlichkeit der Rückmeldungen überprüft.

4 Fragestellungen und Hypothesen

Die erste und zentrale Fragestellung dieses Experiments bezieht sich auf die Feedbackeffekte hinsichtlich der Leistung:

1. Welche Auswirkungen hat Feedback auf das Textverständnis bzw. die Lesekompetenz?

Es wird vermutet, dass die elaborierten Feedbackarten Fehlererklärung, metakognitiver Prompt und Inferenzprompt im Vergleich zu den beiden Kontrollbedingungen (Knowledge of Result, kein Feedback) einen positiven Effekt auf die Leistung bewirken. Dabei wird aufgrund theoretischer Überlegungen zu den unterschiedlichen kognitiven Anforderungen der drei elaborierten Feedbacks in Hinblick auf das Textverstehen (vgl. Abschnitt 3.5.3) angenommen, dass das Feedback Inferenzprompt die höchste Leistungssteigerung nach sich zieht. Dieses Feedback bietet die konkreten kognitiven Schritte, die zum Herstellen der in den Testaufgaben geforderten Inferenzen notwendig sind. Damit sollten diese Rückmeldungen defizitäre, zentrale Prozesse des Textverstehens auffangen, Strategien zur Bewältigung von Verständnisschwierigkeiten bzw. zum Aufbau einer kohärenten, elaborierten Textrepräsentation anbieten und damit dem Herstellen von Inferenzen unmittelbar und nachhaltig dienlich sein.

Die beiden anderen Feedbackarten, Fehlererklärung und metakognitiver Prompt, sind dagegen weniger spezifisch und erfordern stärker den selbstregulierten Einsatz von angemessenen Such- und Korrekturprozessen. Da im Allgemeinen eher Schwierigkeiten in der Ausführung dieser Prozesse vorliegen sollten (das Feedback wird nach Fehlern gegeben), ist davon auszugehen, dass ein Prompt zur Anregung der Verstehensüberwachung oder die Erklärung des gemachten Fehlers weniger zuverlässig und erfolgversprechend zur korrekten Ausführung der notwendigen Prozesse führt. Der Lerneffekt aus dem metakognitiven Prompt und der Fehlererklärung sollte insgesamt kleiner ausfallen als bei den Inferenzprompts. Es wird weiterhin erwartet, dass sich die positiven Effekte der drei elaborierten Feedbackarten und dabei die Überlegung der Inferenzprompts sowohl in der Leistung im Treatment als auch im unmittelbaren Posttest und Follow-up zeigen.

Die Annahmen bezüglich der Auswirkungen auf die Leistung lassen sich wie folgt zusammenfassen:

a) Die Feedbackbedingung Inferenzprompt führt zu einer höheren Leistung als alle anderen Bedingungen.
b) Die Feedbackbedingungen Fehlererklärung und Metakognitiver Prompt bewirken eine höhere Leistung als die Kontrollbedingungen Knowledge of Result sowie die feedbackfreie Gruppe.
c) Die Bedingung Knowledge of Result und die feedbackfreie Kontrollgruppe unterscheiden sich hinsichtlich ihrer Leistungen nicht voneinander.

Die zweite Fragestellung bezieht sich auf die Auswirkungen der Feedbackinterventionen auf die Bearbeitungszeiten:

2. Führt die Intervention mit Feedback zu einer länger andauernden Bearbeitung der Testaufgaben?

Auch wenn in der Feedbackliteratur keine Untersuchungen vorliegen, die sich dieser Frage gewidmet hätten, wird vermutet, dass die elaborierten Feedbackinterventionen während des Treatments insgesamt mehr Zeit für die Beantwortung von Aufgaben aufbringen als die feedbackfreie Kontrollgruppe und die Gruppe mit Knowledge of Result. Diese Vermutung ist an die erwarteten Effekte der elaborierten Feedbacks bezüglich der Leistung geknüpft. Die Rezeption und Umsetzung der elaborierten Rückmeldungen ist ressourcenfordernd und daraus muss im Allgemeinen eine Erhöhung der Verweilzeiten bei den Antworten im Treatment resultieren. Dies betrifft nicht nur die Bearbeitung einer Aufgabe, nachdem für diese das elaborierte Feedback gegeben wurde (Zweitantworten), sondern der Effekt sollte sich auch auf die Beantwortung der Erstantworten im Treatment übertragen.

Ob sich die Kontrollbedingung Knowledge of Result darüber hinaus in der Bearbeitungszeit vom der feedbackfreien, konventionellen Testbedingung unterscheiden wird, kann nicht vorhergesagt werden. Aus theoretischer Perspektive erscheint es sowohl plausibel, dass auch das Feedback Knowledge of Result mehr Such- und Korrekturprozesse beim Lesen auslöst, als auch, dass aufgrund der fehlenden inhaltlichen Hilfestellung für den Lerner im Allgemeinen keine Aussicht auf Erfolg besteht, weshalb er auch eher keine weiteren, zeitfordernden Suchprozesse vornimmt.

Die dritte Fragestellung bezieht sich auf die Auswirkungen der Feedbackinterventionen auf die Testangst:

3. Führt die Intervention mit Feedback zu einer Erhöhung der Testangst?

Wie in Abschnitt 3.5.2 beschrieben gehört Testangst zu den Personmerkmalen, die die Leistung und auch die Feedbacknutzung nachteilig beeinflussen können. Die Feedbackstudien mit Berücksichtigung der Testangst bzw. Ängstlichkeit (vgl. Abschnitt 3.5.2) sprechen hier für einen negativen Einfluss auf die Leistung (Hansen, 1974) und für eine ablehnende Haltung leistungsängstlicher Personen gegenüber Leistungsrückmeldungen (Frey et al., 1986). In dieser Untersuchung sollen auch die Auswirkungen der Feedbackinterventionen auf das situationsspezifische Merkmal Testangst untersucht werden. Es stellt sich die Frage, ob die Feedbackinterventionen aufgrund ihrer vermutlich hohen Anforderungen und möglichen Belastungsmomente (wiederholte Fehlerrückmeldungen etc.) zu einer Verschlechterung im Sinne höherer Testangstwerte führen, die die mögliche Wirkung von Feedback untergraben könnten.
Diesbezüglich wird vermutet, dass im Allgemeinen keine signifikant negativen Effekte der Feedbackintervention zu erwarten sind. Der Kontext des Experiment (d.h. Klassenverband, Schule statt Labor) und insbesondere die low-stake Testbedingung lassen vermuten, dass im Allgemeinen keine nachteiligen Auswirkungen der Untersuchung auf das Befinden der Teilnehmer festzustellen ist. Zudem wird angenommen, dass durch die ausschließlich computerbasierte Intervention (d.h. keine Beurteilung durch eine Person) und die inhaltliche Ausrichtung des Treatments bei den elaborierten Feedbacks (d.h. nicht nur Fehlerrückmeldungen, sondern auch Hilfestellungen) einem möglichen Anstieg von Besorgnis und Aufgeregtheit bezüglich des eigenen Abschneidens entgegengewirkt wird.

Die vierte Fragestellung bezieht sich auf die Wahrnehmung der Rückmeldungen durch die Lerner:

4. Werden die Feedbackinterventionen hinsichtlich ihrer Nützlichkeit unterschiedlich wahrgenommen?

Mit Bezug auf die erwarteten Effekte der elaborierten Feedbackinterventionen auf die Leistung wird vermutet, dass die Einschätzungen der Probanden, die Inferenzprompts erhielten, positiver ausfällt als die Einschätzungen der Probanden, die mit dem metakognitiven Prompt oder den Fehlererklärungen arbeiteten. Die Beurteilung der Intervention mit Knowledge of Result fällt vermutlich am schlechtesten aus.

5 Methodik

5.1 Untersuchungsdesign

Die Untersuchung bestand aus drei Sitzungen: 1) eine Vortestung zur Erfassung relevanter Hintergrund- bzw. Kontrollvariablen, 2) das Experiment mit unmittelbar anschließendem Posttest und 3) ein Follow-up.

Dem Experiment liegt ein einfaktorielles, *between-subjects* Design mit dem fünffach gestuften Faktor Feedbackart zugrunde:

- *Experimentalgruppe: „Knowledge of Result":*
  Das Feedback besteht in der Mitteilung, dass die gewählte Antwort falsch ist – „Das ist falsch".

- *Experimentalgruppe: „Metakognitiver Prompt":*
  Das ist „Knowledge of Result" kombiniert mit dem generellen Hinweis zur Überwachung bzw. Kontrolle des Verständnisses – „Das ist falsch. Prüfe noch einmal, wie der Text richtig zu verstehen ist".
  Es handelt sich hier für alle Items um dieselbe Mitteilung.

- *Experimentalgruppe: „Fehlererklärung":*
  Das ist „Knowledge of Result" kombiniert mit einer Erklärung, warum die gewählte Antwort falsch ist. Die Fehlererklärung kann darauf bezogen sein, dass sich die Aussage des Distraktors nicht mit dem Text vereinbaren lässt, weil die Informationen nicht oder anders im Text enthalten sind (z.B. „Das ist falsch. Die Vorratslager werden tatsächlich im Sommer angelegt."). Zudem kann das Feedback erklären, warum zwei Ereignisse nicht zusammenhängen, die Ursache-Wirkungs- Folge oder die zeitliche Abfolge von Ereignissen eine andere ist (z.B. „Das ist falsch. Wenn Sommer kühler und Winter wärmer werden, wird der Temperaturunterschied zwischen beiden Jahreszeiten geringer.", „Das ist falsch. Der Kapitän ist anfangs in seiner Kajüte und kommt erst später raus.", „Das ist falsch. Im Text steht, dass er schrie, als wenn ihn etwas würgte. Das ist aber nur ein Vergleich, mit dem beschrieben wird, wie sich das Schreien anhört.").
  Die Rückmeldungen sind auf jeden Distraktor eines jeden Items angepasst.

- *Experimentalgruppe: „Inferenzprompt“:*
  Das ist „Knowledge of Result” kombiniert mit einem Hinweis, wie die durch die Fragestellung geforderte Inferenz gezogen werden kann. Dabei wird aufgefordert, den Zusammenhang zwischen Ereignissen herzustellen, eine zeitliche Abfolge von Ereignissen zu berücksichtigen, die Ursache einer Situation und/oder die Wirkung eines Ereignisses zu überprüfen bzw. einzubeziehen (z.B. „Das ist falsch. Finde heraus, in welcher Stimmung sich der Kapitän beim Schrei befindet und durch welches vorige Ereignis diese am besten erklärt werden kann.“, „Das ist falsch. Überprüfe, was sich für die beiden Mannschaften ändert, während sie in den Hauptlagern warten.“, „Das ist falsch. Überprüfe, wann diese angespannte Stimmung aufkommt und überlege, welches vorherige Ereignis eine sinnvolle Erklärung dafür ist.“, „Das ist falsch. Finde heraus, was passiert, bevor der Vater am Anfang der Geschichte weiterfährt.“).
  Die Rückmeldungen sind jeweils auf die Fragestellung eines Items ausgerichtet und damit gibt es in dieser Gruppe eine Mitteilung pro Item.

- *Kontrollgruppe:* kein Feedback

Die Zuordnung der Untersuchungsteilnehmer zu den Versuchsbedingungen erfolgte innerhalb jeder untersuchten Schulklasse randomisiert. Als abhängige Variable wird das Textverständnis/die Lesekompetenz zu verschiedenen Zeitpunkten beobachtet: im Experiment selbst, im Posttest und im Follow-up.

5.2 Stichprobe

An der Untersuchung nahmen insgesamt 664 Schüler der sechsten Klassenstufe teil. Die Gesamtstichprobe rekrutierte sich aus acht Hauptschulen, drei Realschulen und drei Gymnasien aus dem Regierungsbezirk Oberfranken (Bayern). Dabei besuchten 36,0 % eine Hauptschule, 35.2 % eine Realschule und 28.8 % ein Gymnasium. Die Stichprobe bestand aus 334 Mädchen (50.3 %) und 330 Jungen (49.7 %). Das durchschnittliche Alter der Probanden betrug in der ersten Sitzung 12 Jahre und 1 Monat (*SD* = 0;7 Jahre; für die weiteren Sitzungen siehe Tabelle 5). Um einen Indikator für den Migrationshintergrund der Probanden zu erhalten, wurden sie gefragt, welche Sprache sie hauptsächlich zu Hause sprechen. Darauf haben insgesamt 44 Teilnehmer der Untersuchung (6.8 %) eine nicht deutsche Sprache angegeben.

Die Teilnahme an der Untersuchung war zunächst den Schulen und dann auch allen Schülern bzw. deren Erziehungsberechtigten freigestellt. Die Untersuchung war vom Bayerischen Staatsministerium für Unterricht und Kultus (für Realschulen und Gymnasien) sowie der Regierung von Oberfranken (für Hauptschulen) genehmigt worden.

Nicht alle Schüler waren zu allen drei Untersuchungssitzungen anwesend. Die Stichprobengrößen der einzelnen Untersuchungszeitpunkte weichen deshalb von der Gesamtstichprobe von N = 664 Schülern ab. In Tabelle 5 sind die Stichprobenumfänge der einzelnen Sitzungen und deren Eigenschaften zusammengefasst.

Tabelle 5 Überblick über die Teilstichproben

		1. Sitzung Vortestung	2. Sitzung Experiment mit Posttest	3. Sitzung Follow-up
N		609	566	609
Geschlecht	Mädchen (%)	309 (50.7)	300 (53.0)	308 (50.6)
	Jungen (%)	300 (49.3)	266 (47.0)	301 (49.4)
Sprache [a]	Deutsch (%)	568 (93.3)	526 (92.9)	547 (93.0)[b]
	nicht deutsch (%)	41 (6.7)	40 (7.1)	41 (7.0)[b]
Schulart	Hauptschule (%)	223 (36.6)	205 (36.2)	213 (35.0)
	Realschule (%)	203 (33.3)	189 (33.4)	215 (35.3)
	Gymnasium (%)	183 (30.0)	172 (30.4)	181 (29.7)
Alter (in Jahren und Monaten)	*M* *(SD)*	12;1 Jahre [c] (0;7 Jahre)	12;2 Jahre (0;10 Jahre)	12;4 Jahre [d] (0;9 Jahre)

Anmerkungen. [a] Grundlage ist hier die Angabe, welche Sprache hauptsächlich zu Hause gesprochen wird.
[b] Angaben von 588 Probanden.
[c] Angaben von 603 Probanden.
[d] Angaben von 608 Probanden.

5.3 Instrumente

Im Rahmen der Untersuchung wurden eine Reihe kognitiver und motivational-emotionaler Personmerkmale erfasst, die als wichtige Einflussvariablen auf das Textverstehen/die Lesekompetenz und/oder die Bearbeitung des (feedbackgestützten) Tests gelten: Lesegeschwindigkeit, kognitive Grundfähigkeiten (figurale Intelligenz), Testangst und als Komponenten der Lesemotivation das Leseinteresse, Selbstwirksamkeit und Zielorientierungen (vgl. Artelt, Schiefele et al., 2001; Möller & Schiefele, 2004; Rost, D. H. & Schilling, 2006; Shute, 2008). Sie wurden im Wesentlichen als Kontroll- bzw. Hintergrundvariablen erfasst, um somit die Versuchsgruppen, zusätzlich zur durchgeführten Randomisierung, im Vorfeld der Leistungsanalysen auf die Gleichverteilung hinsichtlich dieser Maße prüfen zu können.

Die Erfassung der Testangst dient zusätzlich der Beantwortung der Fragestellung nach Auswirkungen der Feedbackinterventionen auf das Befinden der Probanden. Zudem wurde die subjektiv wahrgenommene Nützlichkeit der Rückmeldungen erfragt, ebenfalls um hier Effekte des Feedbacks überprüfen zu können.

Die Instrumente zur Erfassung der Testangst und der Nützlichkeit der Hilfen wurden im Rahmen der computerbasierten Testung in der zweiten Untersuchungssitzung administriert. Die Leistungstests (Lesegeschwindigkeit und figurale Intelligenz) und die Fragebogen zu ausgewählten Komponenten der Lesemotivation (Leseinteresse, Zielorientierung, Selbstwirksamkeitserwartungen) wurden mittels Papier-und-Bleistift-basierter Verfahren in der Vortestung erhoben.

In der nachfolgenden Beschreibung der Instrumente sind jeweils auch die in der vorliegenden Untersuchung empirisch gefundenen Kennwerte angegeben. Für die Instrumente der Vortestung sind dabei als Auswertungsstichprobe die Schüler herangezogen, die sowohl an der Vortestung als auch am Experiment teilgenommen haben (N = 530[3]). Die empirische Beschreibung der Instrumente, die in der Sitzung des Experiments administriert wurden, basiert auf den Daten der Schüler, die am Experiment, aber nicht notwendigerweise an der Vortestung teilgenommen haben (N = 566).

[3] Der Stichprobenumfang für die Menge der Schüler, die sowohl an der Vortestung als auch am Experiment teilgenommen haben, beträgt insgesamt zwar N = 532. Aber in zwei Fällen kann das Testheft wegen instruktionswidrigen Verhaltens bei der Testbearbeitung nicht ausgewertet werden (vgl. Abschnitt 5.8), so dass sich hier eine Auswertungsstichprobe von N = 530 ergibt.

5.3.1 Leistungstests

Lesegeschwindigkeit

(administriert in Vortestung)

Die Lesegeschwindigkeit gilt als ein guter Indikator der basalen Lesefertigkeit (Rosebrock & Nix, 2006), die die Grundlage für die Prozesse auf Satz- und Textebene bildet (Richter & Christmann, 2002). Eine geringe Lesefertigkeit verweist auf eine nur geringe Automatisierung der Dekodierprozesse, die wiederum kognitive Ressourcen für die Leseprozesse auf hierarchiehöheren Ebenen bindet. Zudem kann vermutet werden, dass sich geringe Verarbeitungskapazitäten aufgrund einer defizitären Lesefertigkeit auch auf die Verarbeitung von Feedback auswirken. Um die Versuchsgruppen zumindest auf die Gleichverteilung hinsichtlich der basalen Lesefertigkeit bzw. Lesegeschwindigkeit überprüfen zu können, wurde dieses Konstrukt als Kontrollvariable erhoben. Das Erfassen der Lesegeschwindigkeit als Indikator der Lesefertigkeit ermöglicht zudem den Einsatz eines ökonomisch durchzuführenden Papier-und-Bleistift Verfahrens – das „Salzburger Lesescreening für die Klassen 5 bis 8“ (SLS 5-8; Auer, Gruber, Mayringer & Wimmer, 2005).

Das Salzburger-Lesescreening ist ein Speedtest. Er besteht aus 70 einfachen Sätzen, die semantisch richtig oder falsch sind (z.B. „Im Mathematikunterricht arbeitet man viel mit Zahlen.“, „Jemand, der Bücher schreibt, ist ein Sänger.“). Die Aufgabe besteht darin, innerhalb von drei Minuten für möglichst viele Sätze zu entscheiden, ob sie wahr sind oder nicht. Die Anzahl der richtig eingeschätzten Sätze ergibt den Gesamttestwert, der theoretisch Werte von 0 bis 70 annehmen kann.

Tabelle 6 Kennwerte des Lesegeschwindigkeitstests

	Lesegeschwindigkeit (theoretisches Max: 70)
N	529
Mittelwert	35.23
Standardabweichung	7.92
Minimum	17
Maximum	68

Die Kennwerte für die vorliegende Stichprobe sind in Tabelle 6 zusammengefasst. Die empirisch gefundenen Werte (M = 35.23, SD = 7.92) sind mit denen der Normstichprobe für sechste Klassen (M = 35.4, SD = 7.8) vergleichbar. Ein Reliabilitätsmaß kann für die vorliegende Stichprobe nicht berechnet werden. Dem Manual des Salzburger Lesescreenings ist jedoch eine Paralleltest-Reliabilität von .89 (Auer et al., 2005, S. 9) zu entnehmen, so dass das Verfahren als sehr reliabel einzustufen ist.

Figurale Intelligenz

(administriert in Vortestung)

Kognitive Grundfähigkeiten gehören zu den basalen Fähigkeiten, die die Lesekompetenz (Rost, D. H. & Schilling, 2006) und auch die Verarbeitung von Feedback (Shute, 2008) beeinflussen. Zur Erfassung wurde ein sprachfreies Maß gewählt, das über den Subtest Figurenanalogien aus dem „Kognitiven Fähigkeitstest“ (KFT 4-12+R; Form A) von Heller und Perleth (2000) erhoben wurde. Der Untertest Figurenanalogien besteht aus 25 Items. Dabei muss jeweils das Verhältnis der Merkmalseigenschaften eines vorgegebenen Figurenpaares erschlossen und auf ein neues Figurenpaar übertragen werden. Vom neuen Figurenpaar ist dabei nur die erste Figur vorgegeben, das passende Gegenstück muss aus fünf Alternativen ausgewählt werden. Die Summe der richtig beantworteten Items ergibt den Gesamttestwert, der theoretisch von 0 bis 25 reichen kann. Diese Spannbreite findet sich auch in der untersuchten Stichprobe (vgl. Tabelle 7). Die interne Konsistenz des Subtests Figurenanalogien ist in dieser Untersuchung mit α = .90 (Cronbachs α) als sehr gut zu bewerten.

Tabelle 7 Kennwerte der Skala Figurale Intelligenz

	Figurale Intelligenz (theoretisches Max: 25)
N	530
Mittelwert	16.40
Standardabweichung	6.30
Minimum	0
Maximum	25
Cronbachs α	.90

5.3.2 Fragebogen

Leseinteresse

(administriert in Vortestung)

Der Begriff Leseinteresse bezieht sich auf das spezifische Interesse am Lesen. Das Leseinteresse beeinflusst das Leseverhalten und steht im positiven Zusammenhang mit dem Textverstehen bzw. die Lesekompetenz (Guthrie & Wigfield, 2000). Zur Erfassung des Leseinteresses wurden neun Items eingesetzt, die dem „Berliner Leselängsschnitt“ (McElvany, Kortenbruck & Becker, 2008, Skala Lesemotivation) und „PISA 2000“ (Kunter et al., 2002, Skala Leselust) entnommen sind (siehe Anhang A). Die Items beziehen sich auf die Dimensionen der individuellen Bewertung von Inhaltsbereichen, positiver emotionaler Erfahrungen im Zusammenhang mit dem Lesen und der Selbstintentionalität.

Das Antwortformat ist eine vierstufige Skala von 1 für „stimmt gar nicht“ bis 4 für „stimmt völlig“. Die interne Konsistenz der Skala ist in dieser Untersuchung mit einem $\alpha = .93$ (Cronbachs α) als sehr gut zu beurteilen (vgl. für weitere Kennwerte Tabelle 8).

Tabelle 8 Kennwerte der Leseinteresseskala

	Leseinteresse (9 Items)
N	530
Mittelwert	2.76
Standardabweichung	0.80
Cronbachs α	.93

Selbstwirksamkeit

(administriert in Vortestung)

Das Konstrukt der Selbstwirksamkeit bezieht sich auf die Erwartungen einer Person, selbst in der Lage zu sein, ein bestimmtes Verhalten erfolgreich auszuführen. Nach Bandura (1977) werden Selbstwirksamkeitserwartungen insbesondere dann handlungsleitend, wenn etwa bei der Bearbeitung einer Aufgabenstellung Schwierigkeiten auftreten. Selbstwirksamkeit kann „domänenspezifisch bzw. fachspezifisch und/oder aufgabenspezifisch erfasst und auch zu konkreten Leistungen in spezifischen Aufgaben in Beziehung gesetzt (werden)“ (Köller & Möller, 2006, S. 696).

In dieser Untersuchung wurden zur Erfassung von Selbstwirksamkeitserwartungen insgesamt 11 Items verwendet, die PISA 2000 (Kunter et al., 2002, Skala Self-Efficacy) und dem Berliner Leselängsschnitt (McElvany et al., 2008, Skala Selbstwirksamkeit Lesen) entnommen wurden (siehe Anhang A).

Das Antwortformat ist eine vierstufige Skala von 1 für „stimmt gar nicht“ bis 4 für „stimmt völlig“. Die für diese Stichprobe ermittelten Kennwerte der Skala sind in Tabelle 9 zusammengefasst. Die interne Konsistenz beträgt $\alpha = .80$ (Cronbachs α) und ist damit als gut zu bewerten.

Tabelle 9 Kennwerte der Selbstwirksamkeitsskala

	Selbstwirksamkeits-erwartungen (11 Items)
N	519
Mittelwert	3.05
Standardabweichung	0.70
Cronbachs α	0.80

Zielorientierungen

(administriert in Vortestung)

Mit dem Konstrukt der Zielorientierungen sind motivationale Tendenzen bzw. Personmerkmale umschrieben, die die situationsspezifische Lernmotivation beeinflussen (Köller & Baumert, 1998). Dabei werden häufig zwei Zielorientierungen gegenübergestellt, die mit unterschiedlichen Begriffspaaren belegt sind (Köller & Baumert, 1998). Eine Unterteilung ist die in Aufgaben- und Ichorientierung sensu Nicholls (1984)[4]. Im Zusammenhang mit Feedback sind die Zielorientierungen deshalb relevant, weil damit auch Annahmen über den Umgang mit Misserfolg in Leistungssituationen getroffen werden können. Personen mit Aufgabenorientierung reagieren auf Misserfolge oder Schwierigkeiten eher mit bewältigendem Verhalten, da sie eher davon überzeugt sind, dass eine zusätzliche Anstrengung zur Bewältigung von Schwierigkeiten führen und zum Erwerb neuer Kompetenzen beitragen kann. Dagegen tendieren Personen mit Ichorientierung dazu, bewältigendes Verhalten nach Misserfolg

[4] Nicholls (1984) verwendet in dem englischsprachigen Original die Begriffe „*task involvement*“ und „*ego involvement*“.

nur dann einzusetzen, wenn sie sich als hoch kompetent einschätzen. Bei Personen mit Ichorientierung, die sich in dem relevanten Fähigkeitsbereich als nicht kompetent sehen, führen Misserfolge dann eher zu hilflosem Verhalten und ein Lernfortschritt ist nicht zu erwarten (Köller & Baumert, 1998).

Zur Erfassung der Zielorientierungen wurden für die vorliegenden Untersuchung insgesamt neun Items der Skalen Aufgabenorientierung und soziale Vergleichsorientierung aus der Studie „Bildungsverläufe und psychosoziale Entwicklung im Jugendalter“ (BIJU; Baumert, Gruehn, Heyn, Köller & Schnabel, 1997) genutzt (siehe Anhang A). Für fünf der 11 Items wurden die sprachlich leicht angepassten Formulierungen aus PISA 2000 (Kunter et al., 2002, Skalen Task Orientation und Ego Orientation) verwendet. Fünf der neun eingesetzten Items beziehen sich auf die Aufgabenorientierung, vier Items repräsentieren die Ichorientierung.

Tabelle 10 Kennwerte der Skalen zu Zielorientierungen

	Ich-orientierung (4 Items)	Aufgaben-orientierung (5 Items)
N	530	530
Mittelwert	2.73	3.15
Standardabweichung	0.80	0.58
Cronbachs α	.82	.76

Alle Items sind wiederum auf einer vierstufigen Skala von 1 für „stimmt gar nicht“ bis 4 für „stimmt völlig“ zu beantworten. Die für die vorliegende Untersuchung gefundenen empirischen Kennwerte der Skalen sind in Tabelle 10 dargestellt. Die interne Konsistenz der Skalen ist mit α = .82 bzw. α = .76 (beides Cronbachs α) als gut bzw. akzeptabel zu bewerten. Beide Skalen korrelieren mit r = .28 (p < .001) und weisen damit einen eher geringen Zusammenhang auf. Beide Skalen werden als separate Maße und nicht als eine zusammengefasste Skala in die Auswertungen einbezogen werden.

Testangst

(administriert im Experiment – vor und nach der Treatmentphase)

Das Konstrukt der Testangst (oder Prüfungsängstlichkeit) bezieht sich auf Befürchtungen und damit zusammenhängende Reaktionen auf unterschiedlichen Ebenen bezüglich möglicher Fehler oder negativer Konsequenzen in einer Leistungs- bzw. Bewertungssituation (Zeidner, 2007). In der Erfassung des Konstrukts wird meistens von einem relativ stabilen Persönlichkeitsmerkmal, das aber situationsspezifisch zu erfassen ist, ausgegangen (Spielberger & Vagg, 1995).

Als Instrument zur Erfassung der Testangst wurde auf das „Prüfungsängstlichkeitsinventar TAI-G“ (Hodapp, 1991) zurückgegriffen, aus dem acht der insgesamt 30 Items entnommen wurden. Bei einem Item wurde eine sprachliche Anpassung vorgenommen: aus „Ich fühle mich unbehaglich“ wurde „Ich fühle mich unwohl“. Die ausgewählten Items (siehe Anhang A) sind der kognitiven Komponente der Besorgtheit (N = 5 Items) und der affektiven Komponente der Aufgeregtheit (N = 3 Items) zuzuordnen. Besorgtheit und Aufgeregtheit gelten als die grundlegenden Komponenten von Testangst (Hodapp, 1991), die zwar in verschiedenen Untersuchungen mit teilweise leicht unterschiedlichen Skalenkonzeptionen mehrheitlich hoch miteinander korrelieren, aber aus empirischer Sicht dennoch als separate Subskalen geführt werden (Zeidner, 2007).

Der Fragebogen zur Erfassung der Testangst wurde unmittelbar vor und nach dem Experiment computerbasiert erfasst. Die Items, die vor dem Experiment administriert wurden, waren darauf ausgerichtet, was die Probanden empfinden, wenn sie an die bevorstehende Testbearbeitung denken (z.B. „Ich denke daran, was passiert, wenn ich schlecht abschneide.“). Die Items, die unmittelbar nach dem Experiment vorgelegt wurden, erfragten das Empfinden bezüglich der zurückliegenden Bearbeitung der Texte und Aufgaben (z.B. „Ich denke daran, was passiert, wenn ich schlecht abgeschnitten habe.“). Diese retrospektive Variante ist im Originalfragebogen von Hodapp (1991) nicht vorgesehen und stellt damit eine entsprechende sprachliche Adaption der Originalitems dar (siehe Anhang A). Das Antwortformat war für beide Varianten eine vierstufige Skala von 1 für „stimmt gar nicht“ bis 4 für „stimmt völlig“.

Tabelle 11 Kennwerte der Subskalen zur Testangst

	Testangst			
	Subskala Besorgtheit (5 Items)		Subskala Aufgeregtheit (3 Items)	
	Prä	Post	Prä	Post
N	566	561	566	561
Mittelwert	2.29	2.29	1.64	1.69
Standardabweichung	0.69	0.81	0.62	0.75
Cronbachs α	.80	.86	.62	.76

Die Auswertung erfolgt in Anlehnung an Hodapp (1991) für beide Komponenten bzw. Subskalen, die Aufgeregtheit und die Besorgtheit, getrennt. Die entsprechenden Kennwerte sind in Tabelle 11 zusammengefasst. Die internen Konsistenzen für die Subskala Besorgtheit sind mit Cronbachs α = .80 und α = .86 als gut zu bewerten. Die interne Konsistenz der Subskala Aufgeregtheit beträgt in der Prätestversion lediglich α = .62 und in der Posttestversion α = .76. Die zwei Subskalen korrelieren zu r = .56 (p < .001) in der Prätestversion und zu r = .64 (p < .001) in der Posttestversion.

Einschätzung zur Nützlichkeit der Feedbacks

(administriert nach Treatmentphase)

Unmittelbar nach dem Experiment werden zusätzlich fünf Items computerbasiert dargeboten, die eine Einschätzung der subjektiv wahrgenommenen Nützlichkeit der angebotenen Rückmeldungen erfordern. Da die Kontrollgruppe ohne Feedbackintervention arbeitet, treten diese Items nur in den Computerprogrammen der Experimentalgruppen auf. Die Items sind eine Eigenentwicklung und orientieren sich an inhaltlichen Gesichtspunkten und fragen nach unterschiedlichen Aspekten, die für die Diskussion oder die Weiterentwicklung der verschiedenen Feedbackinterventionen relevant sein können. Die Items sind die folgenden:

„Bitte schätze ein, wie du die Rückmeldungen in den Texten fandest!

1. Ich fand sie hilfreich.
2. Ich fand sie verwirrend.
3. Sie haben mich abgelenkt.
4. Sie haben mir geholfen, die Aufgabe doch noch zu bewältigen.
5. Sie haben mir bei darauffolgenden Aufgaben geholfen.“

Die Beantwortung der Items ist wiederum auf einer vierstufigen Skala von 1 für „stimmt gar nicht“ bis 4 für „stimmt völlig“ vorzunehmen. Die Kennwerte der Skala sind in Tabelle 12 zusammengestellt. Die interne Konsistenz fällt mit einem α = .63 (Cronbachs α) äußerst niedrig aus.

Tabelle 12 Kennwerte der Skala zur wahrgenommenen Nützlichkeit der Feedbacks

	Nützlichkeit der Feedbacks (5 Items)
N	452 [a]
Mittelwert	2.79
Standardabweichung	0.60
Cronbachs α	.63

Anmerkungen. [a] Die Auswertungsstichprobe umfasst nur die Schüler der experimentellen Bedingungen, die Kontrollgruppe ist hier ausgeschlossen.

5.4 Material (Texte und Items)

Die Texte und Items, die in dieser Untersuchung zur Erfassung des Textverständnisses/der Lesekompetenz eingesetzt wurden, unterlagen bestimmten Anforderungen. Diese ergeben sich zum einen aus dem inhaltlichen Schwerpunkt, der auf den hierarchiehöheren Prozessen des Textverstehens lag, zum anderen aus praktischen Notwendigkeiten (z.B. eine gewisse Anzahl an Antwortalternativen pro Item). Da entsprechende Testverfahren/-materialien für die Zielgruppe der Sechstklässler nicht bestanden bzw. zugänglich waren, wurden die Texte und Items für diese Untersuchungen selbst entworfen[5].

[5] Bei Bedarf sind die Texte und Aufgaben bei der Autorin anzufordern.

5.4.1 Konzeption der Texte und Items

Beschreibung des Textmaterials

Kriterien für die Textauswahl: Die verwendeten Texte sind zu fast gleichen Teilen Sach- und Erzähltexte. Sie wurden älteren Lehr- oder Jugendbüchern entnommen oder aus Texten verschiedener Quellen (z.B. Bücher, Onlineartikel, Zeitungsartikel) zu einem Thema zusammengestellt. Alle Texte wurden dahingehend überprüft und angepasst, dass sie mit Blick auf die Altersgruppe der Probanden verständlich sind. Fremdwörter und schwierige Satzkonstellationen wurden vermieden.

Die Texte behandeln Themen der Biologie, Geschichte oder Geografie oder sind Auszüge fiktionaler Geschichten. Bei der Auswahl der Textinhalte wurde berücksichtigt, dass sie weder einen Bias hinsichtlich Geschlecht, Kultur oder Religion noch emotional belastende Themen beinhalten. Außerdem wurde darauf geachtet, dass kein spezifisches Vorwissen zum Verständnis der Texte notwendig war. Ebenso wurden unterrichtsrelevante Themen (d.h. aktueller Lehrstoff) der Zielgruppe bzw. aktuelle Lehrbuchtexte ausgeschlossen. Zu diesem Zweck wurde einerseits eine Recherche der für Bayern zu der Zeit empfohlenen und zugelassenen Lehrbücher der Fächer Deutsch, Biologie und Geografie/Sachunterricht durchgeführt. Andererseits wurde für die Auswahl der Texte bewusst auch auf ältere Lehrbücher für die sechste Klassenstufe (d.h. Bücher, die zu der Zeit sehr wahrscheinlich nicht mehr im Unterricht für diese Klassenstufe eingesetzt wurden) oder aktuelle Lehrbücher höherer Klassenstufen zurückgegriffen.

Darüber hinaus war es von zentraler Bedeutung, dass die Texte eine ausreichende „Informationsdichte“ boten. Damit ist gemeint, dass ein Text relativ viele Ansatzpunkte für die Konstruktion von Items bieten musste. Denn es sollten pro Text immer mehrere Fragen formuliert werden, die zudem vor allem das tiefere Verständnis des Textes abprüfen sollten. Im Gegensatz zu einer Anforderung, wie beispielsweise das Entnehmen einer einzelnen, explizit genannten Information, bietet ein Text für das Erfragen des tieferen Verständnisses normalerweise nicht außerordentlich viele Anhaltspunkte. Gegebenenfalls wurden, ohne auf die Verständlichkeit der Texte zu verzichten, Originaltexte verdichtet, mit neuen Informationen angereichert oder auch Informationen aus dem Text entnommen, um so Raum für eine ausreichende Anzahl an Items zu schaffen.

Zudem galt es zu berücksichtigen, dass die einzelnen Texte eine gewisse Länge nicht überschreiten. Denn mehrere sehr lange Texte hätten zu einem zu großen Leseaufwand

im Rahmen der Untersuchung geführt. Alle Texte sind zwischen 258 und 430 Wörtern lang. In Tabelle 13 sind die wichtigsten Merkmale der eingesetzten Units zusammengefasst.

Tabelle 13 Charakterisierung des Materials

Unit (Thema)[a]	Textsorte	Textlänge (Wörter)	Anzahl der Items
Experiment			
1. „Durch die Eiswüste der Antarktis“ (Beschrieben wird der Wettlauf zwischen zwei Forschern bei der „Eroberung“ des Südpols.)	Sachtext	300	6
2. „Klimawandel“ (Beschrieben werden Ursachen und Auswirkungen des Klimawandels.)	Sachtext	308	6
3. „Der Sprung“ (Erzählt wird von einem Jungen, der auf einem Schiff mit Mannschaft die Welt umsegelt, dort in eine Auseinandersetzung gerät, ins Meer stürzt und schließlich gerettet werden muss.)	Erzähltext	416	8
4. „Geschichte von Herrn Sommer“ (Erzählt wird von einem Jungen, der mit seinem Vater bei Regen im Auto fährt. Sie treffen auf einen Bekannten und versuchen vergebens ihn davon zu überzeugen, ins Auto einzusteigen.)	Erzähltext	430	9
5. „Nationalpark in Not“ (Berichtet wird von einem realen Nationalpark in Vietnam, der mit finanziellen Schwierigkeiten und Wilderei zu kämpfen hat.)	Sachtext	291	8
Posttest			
1. „Der alte Mann“ (Erzählt wird von einem Jungen und seinem Großvater, der mit Eisfischen versucht, den Lebensunterhalt der Familie zu sichern.)	Erzähltext	349	7
2. „Der Kakaobaum“ (Berichtet wird von Anbaubedingungen der Bäume und wie aus deren Früchten die Kakaobohnen gewonnen werden.)	Sachtext	311	7

Noch Tabelle 13 Charakterisierung des Materials

	Unit (Thema)[a]	Textsorte	Textlänge (Wörter)	Anzahl der Items
	Follow-up			
1.	„Der Elefantenrüsselfisch“ (Berichtet wird vom außergewöhnlichen Aussehen und Orientierungsvermögen dieses Fisches.)	Sachtext	258	6
2.	„Tansania sucht einen eigenen Weg“ (Berichtet wird von der Geschichte Tansanias, seinen Problemen der Armut und Hunger und einem gemeinschaftlichen Lösungsansatz.)	Sachtext	318	7
3.	„Der große Fang“ (Erzählt wird von einem Jungen, der beim Eisfischen auf sich allein gestellt mit einem großen Fisch kämpfen muss.)	Erzähltext	418	9
4.	„Auf dem Bahnhofsvorplatz“ (Erzählt wird von einer Begebenheit, in der ein Afrikaner Soldaten der Kolonialmacht vorführt.)	Erzähltext	422	8

Anmerkungen. [a] Die hier abgebildete Reihenfolge der Units entspricht der „Form A“ der Materialien in den jeweiligen Untersuchungssitzungen. In „Form B“ sind die Units jeweils in umgekehrter Reihenfolge angeordnet.

Beschreibung des Itemmaterials

Anforderungsbereiche der Items: Der Gegenstandsbereich dieser Arbeit sind die hierarchiehöheren Prozesse beim Lesen von Texten. Entsprechend zielen die Items primär auf das Herstellen von Sinnzusammenhängen zwischen benachbarten Satzteilen/Sätzen oder zwischen mehreren Sätzen und/oder über Abschnitte hinweg ab. Die Fragen zum Text beziehen sich typischerweise auf Ursachen von Ereignissen, Ursache-Wirkungs-Gefüge oder Zusammenhänge von Ereignissen. Beispiele sind: „Warum entschließt sich Amundsen zu einer Expedition zum Südpol?“, „Warum werden nach der Ankunft in der Antarktis die Vorratslager in großer Eile angelegt?“, „Was bedeutet es, wenn die Temperaturunterschiede zwischen Sommer und Winter größer werden?“ oder „Warum reißt der Mann vermutlich die Flagge herunter?“.

Darüber hinaus war beabsichtigt, Items zu generieren, die tendenziell eher schwierig für die Zielgruppe sind. Da die Feedbackgabe in dieser Untersuchung an fehlerhafte Antworten geknüpft ist, galt es ausreichend viele Gelegenheiten für eine

Feedbackintervention zu schaffen. Gleichzeitig galten für die Konstruktion der Items allgemeingültige Maßgaben (Rost, J., 2004), beispielsweise dass Distraktoren plausibel gestaltet, Fragestellungen unmissverständlich formuliert und eineindeutig zu beantworten sein müssen.

Antwortformat: Alle Items zu den Texten waren im Multiple-Choice Format gehalten. Dabei gab es pro Item stets fünf Antwortalternativen, wo denen immer nur eine die richtige Antwort war. Die vergleichsweise hohe Anzahl an Antwortalternativen pro Item stellte zwar durchaus Herausforderungen an die Itemkonstruktion. Doch das Experiment war so angelegt, dass beim wiederholten Beantworten eines Items nach der Feedbackgabe automatisch die zuvor gewählte, falsche Antwortalternative entfiel (vgl. Abschnitt 5.5). Auch dann sollte die Ratewahrscheinlichkeit nicht zu hoch ausfallen. Deshalb wurde sich für grundsätzlich fünf Antwortalternativen pro Item entschieden; die Alternative, noch mehr Distraktoren pro Item zu gestalten, wäre nicht für alle Aufgaben verlässlich plausibel umzusetzen gewesen. Das geschlossene Antwortformat wurde gewählt, weil es im Rahmen des computerbasierten Testens das unmittelbare Feedbackgeben eher ermöglicht als zum Beispiel offene Antwortformate.

Bearbeitungsvorgaben: Die Bearbeitung der Texte und Items unterlag keiner Zeitvorgabe. Zur Beantwortung der Items blieb der dazugehörige Text stets einsehbar. Damit stand in der Untersuchung das verstehende Lesen von Texten im Vordergrund. Gedächtnisleistungen, wie sie beim Beantworten von Items ohne Texteinsicht stärker zum Tragen kommen, rückten dagegen in den Hintergrund (vgl. Christmann, 2002). Alle Lesekompetenztests liegen in zwei Formen vor, die sich jeweils dadurch unterscheiden, dass die Reihenfolge der Units umgekehrt ist.
In den computerbasierten Tests (Experiment, Posttest) konnten Items weder übersprungen noch vorherige Items wieder aufgerufen werden (siehe Abschnitt 5.6.1). Im Follow-up, das mithilfe eines Papier-und-Bleistift-basierten Testhefts umgesetzt wurde, war das dagegen möglich.

5.4.2 Kognitive Interviews für die Materialentwicklung

Die Entwicklung des Text- und Itemmaterials wurde durch deren Evaluierung in Kognitiven Interviews begleitet. Kognitive Interviews zählen zu den Verfahren des

Cognitive Laboratory, die im Rahmen der Itementwicklung zur Überprüfung potentieller Probleme bei der Beantwortung von Fragebogenitems oder, wie hier geschehen, Kompetenztestitems eingesetzt werden (OECD, 2005; Prüfer & Rexroth, 2005). Das Ziel von Kognitiven Interviews besteht hauptsächlich darin, das Verständnis der Probanden hinsichtlich des Itemmaterials (Aufgabenstellung, Antwortalternativen etc.) gezielt zu hinterfragen, um daraus Rückschlüsse auf die Qualität bzw. Angemessenheit des Materials ziehen zu können. Das heißt, es wird überprüft, ob die Materialien von der relevanten Stichprobe so verstanden und beantwortet werden (können), wie es bei der Konstruktion beabsichtigt wurde. Die kognitiven Techniken, die dabei sehr häufig angewendet werden, sind das Nachfragen (*Probing*) und das laute Denken (Prüfer & Rexroth, 2005). Das laute Denken hat gegenüber dem Nachfragen jedoch unter anderem den Nachteil, dass Probanden für gewöhnlich einige Eingewöhnungszeit benötigen, bis sie die Methode, das stetige Verbalisieren aller Gedankengänge, hinreichend gut umsetzen (van Someren, Barnard & Sandberg, 1994). Da der Umfang der einzelnen Interviews in der vorliegenden Arbeit vergleichsweise gering gehalten werden sollte (meistens maximal 30 Minuten), wurde hier hauptsächlich auf die Technik des Nachfragens zurückgegriffen.

Im Rahmen dieser Arbeit bestand das Ziel der kognitiven Interviews darin, die Items (Fragestellungen, Antwortalternativen) hinsichtlich Verständlichkeit, Schwierigkeit und Angemessenheit (Verständnis) einzuschätzen sowie die verschiedenen Rückmeldungen in ihrer Wirkung abzuschätzen und die konkreten Mitteilungen ebenfalls hinsichtlich ihrer Verständlichkeit zu prüfen. Die Überprüfung der Verständlichkeit zielte auf das Auffinden missverständlicher Formulierungen oder für die Altersgruppe unbekannte oder ungewöhnliche Wörter/Formulierungen. Dazu wurden Teilnehmer gebeten, fragliche Wörter/Formulierungen zu erklären oder mit anderen Worten wiederzugeben. Ein zentraler Bestandteil der Interviews war das Hinterfragen des Verständnisses der Leser, woraus auch auf die Angemessenheit der Items und Rückmeldungen geschlossen wurde: Damit wurde abgeprüft, ob die Leser die Fragestellungen und Antwortalternativen bzw. die Rückmeldungen in ihren Bedeutungen so verstehen, wie sie bei der Konstruktion gemeint waren. Dabei wurde auch kontrolliert, ob die entworfenen Antworten mit den mentalen Modellen von Lesern sechster Klassen vereinbar und damit angemessen waren. Das bedeutet, es wurde überprüft, ob die als richtig klassifizierten Antworten auch von den etwa zwölfjährigen Lesern als richtige Antworten angesehen wurden und

entsprechend falsche Antworten auch für sie eindeutig falsche Antworten für die betreffende Fragestellung darstellten. Hier spielen Misconceptions und fehlende Erfahrungswelten der Schüler eine Rolle, die eine logisch richtige Antwort für sie dennoch zu einer nicht nachvollziehbaren Lösung werden lassen können. Um das Verständnis und die Angemessenheit von Items und Rückmeldung abzuprüfen, wurden Leser beispielsweise gebeten, zu erklären, warum sie die gewählte Antwort für richtig und die verbliebenden Alternativen für falsch hielten. Ebenso konnten sie aufgefordert werden, die richtige Antwort auf eine Frage frei (Antwortalternativen vorher abgedeckt) zu formulieren, Fragestellungen und/oder Antwortalternativen in eigenen Worten wiederzugeben oder nach dem Lesen eines Textes dessen wesentlichen Inhalt, die wichtigsten Aussagen bzw. Ereignisse zu erzählen. In Bezug auf die Rückmeldungen konnten die Teilnehmer auch gebeten werden, zu formulieren, wie sie zum Beispiel einen Klassenkammeraden, der die entsprechende Testfrage falsch beantworten könnte, unterstützen würden, so dass er zur richtigen Antwort gelangt.

Dem Ablauf nach wurde im Normalfall zunächst ein Text (ohne Zeitvorgabe) vom Probanden leise gelesen. Danach wurden ihm die Items zum Text einzeln vorgelegt, jedes Item sollte zunächst schriftlich beantwortet werden, bevor die Nachfragen zum Item gestellt wurden. Die Feedbacks wurden teilweise vor den Nachfragen zum Item und teilweise danach gegeben und besprochen. Welche Nachfragen konkret und in welcher Reihenfolge gestellt und mit dem Teilnehmer bearbeitet wurden, folgte keinem festen Schema. Sondern der Interviewverlauf wurde diesbezüglich deutlich von den individuellen Teilnehmern und deren Antworten abhängig gemacht (d.h. wenn ein Leser eine Frage richtig beantwortet und begründet hatte, erübrigte sich das Nachfragen nach z.B. Misconceptions, nach einer falschen Testantwort was es hingegen angebracht). Die Interviews fanden in Einzelsitzungen statt. Das Material lag als Papier-und-Bleistift-basierte Variante vor und die Aussagen der Teilnehmer wurden protokolliert.
Die Kognitiven Interviews wurden mit insgesamt 21 Schülern der sechsten Klassenstufe aus den drei relevanten Schulformen im Rahmen der Nachmittagsbetreuung verschiedener Einrichtungen in Bamberg durchgeführt. Die insgesamt 21 Interviews waren über mehrere Wochen verteilt. Um die Belastung für den einzelnen Teilnehmer gering zu halten, bearbeitete jeder maximal zwei Units. Die Teilnehmer wurden über den Zweck der Sitzung (d.h. die Überprüfung des Materials, nicht ihre Leistungsfähigkeit) aufgeklärt.

Die Erkenntnisse zur Qualität und Angemessenheit der Materialien flossen in die Überarbeitung der Materialien ein und die Texte und Items wurden gegebenenfalls wiederholt in Kognitive Interviews gegeben.

5.5 Prozedur der Feedbackgabe

Die Feedbackgabe im Experiment erfolgte nach dem folgenden Prinzip: Wenn eine Aufgabe im ersten Versuch falsch beantwortet wurde, erhielten die Probanden der Experimentalgruppen eine Feedbackmitteilung. Danach war dieselbe Aufgabe ein zweites Mal zu beantworten. Die zuvor gewählte, falsche Antwortalternative konnte dabei nicht mehr ausgewählt werden, war aber noch sichtbar. Nach der zweiten Antwort für ein Item wurde dann automatisch das nächste Item präsentiert. Für Aufgaben, die dagegen im ersten Versuch richtig gelöst wurden, erschien kein Feedback; sie wurden unmittelbar vom nächsten Item gefolgt.
Die Probanden der Kontrollgruppe, die entsprechend dem Versuchsplan nie Feedback erhielten, konnten alle Items nur einmal beantworten. Wenn eine Aufgabe beantwortet war, wurde also automatisch die nächste präsentiert.

5.6 Das computerbasierte Programm

Das computerbasierte Programm, mit dem das Experiment und der sich anschließende Posttest durchgeführt wurden, ist eigens für diese Zwecke programmiert worden. Der Einsatz eines Computerprogramms war für die Gruppentestung notwendig, um die Feedbackgabe zuverlässig und fehlerfrei umsetzen zu können. Daneben ergaben sich daraus weitere Vorteile wie die größtmögliche Standardisierung des Vorgehens, die automatische Speicherung der Daten und die Erfassung der Bearbeitungszeiten. Der Posttest wurde dann aus praktischen Gründen in das Computerprogramm integriert, um nicht innerhalb derselben Sitzung das Medium wechseln zu müssen.

5.6.1 Aufbau

Die computerbasierte Testumgebung setzte sich aus verschiedenen Abschnitten zusammen. Die erste Seite war eine Anmeldemaske, die einige personenbezogene Daten

(Name, Geburtsdatum, Geschlecht, Sprache, Klasse und Schulart) erfragte. Es folgten die Instruktionen zum Ablauf der Sitzung, zum Umgang mit dem Programm und eine Beschreibung der Aufgabe. Darin war auch eine Übungsphase zur Sicherung des Instruktionsverständnisses enthalten. Nach den Instruktionsseiten erschien auf einer Seite der Fragebogen zur Testangst. Die Antwortkategorien waren mit *Radiobuttons* abgebildet. Alle Items mussten eingeschätzt werden, bevor die nächste Seite aufgerufen werden konnte.

Nach dem Fragebogen begann das Experiment. Es erschien der erste Text mit dem ersten dazugehörigen Item. Die grafische Oberfläche bestand immer aus drei Feldern, wie in Abbildung 3 schematisch dargestellt ist. Das Textfeld war scrollbar, da die Länge der Texte die Größe des Fensters überschritt. Die Auswahl einer Antwort erfolgte über das Anklicken des Radiobuttons, der sich vor der gewünschten Antwortalternative befand. Solange der „Weiter"-Button noch nicht gedrückt wurde, konnte die Auswahl einer Antwort beliebig oft korrigiert werden, ohne dass dies aufgezeichnet wurde. In den Experimentalgruppen erschien das Feedback nach falschen ersten Antworten (vgl. Abschnitt 5.5). Dabei blieben der Text und die Aufgabe sichtbar und beim zweiten Antwortversuch war die zuvor ausgewählte, falsche Antwortalternative nicht mehr auswählbar. In der Kontrollgruppe blieb das Feedbackfenster immer leer.

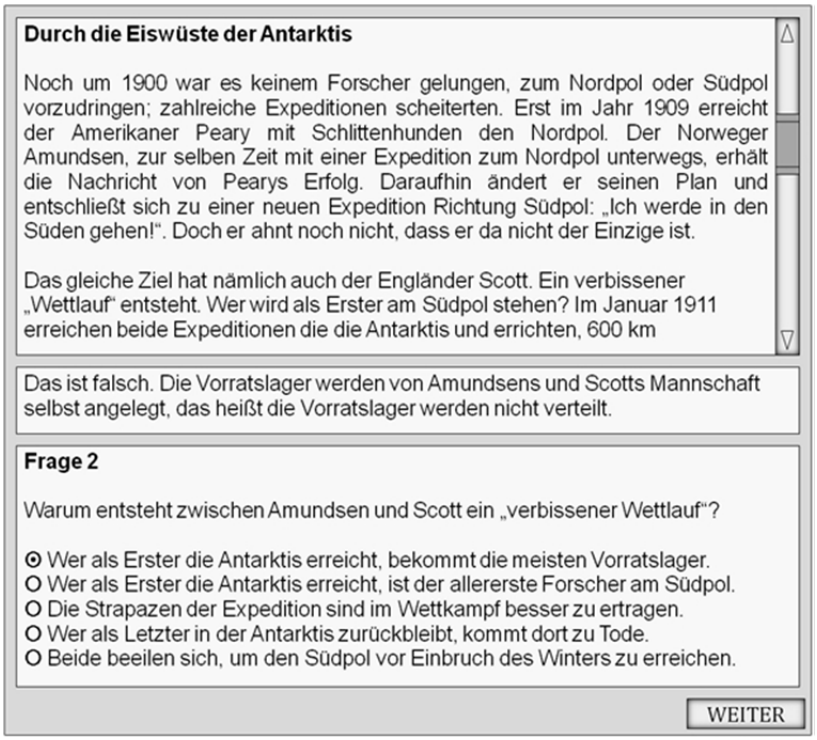

Abbildung 3 Schematische Darstellung der Programmoberfläche im Experiment.

Nachdem alle Items des Experiments bearbeitet waren, erschien erneut ein Fragebogen zur Testangst, in den Experimentalgruppen danach zusätzlich noch die Items zur Einschätzung der Nützlichkeit der Feedbacks. Der Aufbau dieser Seiten und die Beantwortung waren identisch mit dem ersten Fragebogen, der vor dem Experiment administriert wurde.

Im Anschluss an den Testangstfragebogen in der Kontrollgruppe bzw. den Einschätzfragen für die Feedbacks in den Experimentalgruppen begann der letzte Teil des Programms, der Posttest. Die Seiten des Posttest waren analog zu denen des Experiments aufgebaut. Das Feedbackfenster blieb in allen Gruppen stets leer. Das Vorgehen zum Abgeben einer Antwort war ebenfalls wie im Experiment, wobei allen Versuchsgruppen nunmehr nur ein Antwortversuch pro Item zur Verfügung stand.

Die Bearbeitung des Programms unterlag keinen Zeitbegrenzungen. Zudem konnten Items weder übersprungen noch konnte zu vorherigen Aufgaben zurückgegangen werden.

5.6.2 Technische Merkmale

Das erzeugte Computerprogramm basierte auf einem Java Bytecode und konnte damit plattformunabhängig und, gespeichert auf USB-Sticks, ohne Installation auf den Zielrechnern in den Schulen eingesetzt werden. Dies war notwendig, da in den Schulen unterschiedliche technische Bedingungen, verschiedene Betriebssysteme und vermehrt Sperrungen für Installationen und Webanwendungen vorzufinden waren.

Am Ende eines Programmdurchlaufs wurden die Testergebnisse und Bearbeitungszeiten automatisch in einer xml-Ergebnisdatei gespeichert. Das heißt, es wird gespeichert, welche Antwortkategorien in den Fragenbogen und welche Antwortalternativen in den Textaufgaben zu welchem Zeitpunkt ausgewählt wurden. Daneben wurden auch die personenbezogenen Daten aus der Anmeldeseite in die Ergebnisdatei geschrieben. Mithilfe der xml-Ergebnisdateien konnten die Daten in eine Datenbank und darüber in die verwendete Auswertungssoftware SPSS importiert werden.

5.7 Untersuchungsdurchführung

Der Untersuchungsablauf und die Inhalte der drei Untersuchungssitzungen – Vortestung, Experiment mit Posttest und Follow-up – sind in Tabelle 14 zusammengefasst. Die Vortestung diente der Erfassung ausgewählter Hintergrund- bzw. Kontrollvariablen

mittels Papier-und-Bleistift-basierter Verfahren. Diese Sitzung wurde meist von zwei Testleitern geleitet. Den Probanden wurden zu Beginn der Untersuchung Ziel und Zweck der Studie erläutert. Danach instruierte ein Testleiter schrittweise das Vorgehen bei der Bearbeitung der beiden Leistungstest und der vier Fragebogen. Die Sitzung dauerte eine Unterrichtsstunde.

Tabelle 14 Untersuchungsablauf

1. Sitzung (eine Unterrichtsstunde)	2. Sitzung (doppelte Unterrichtsstunde)	3. Sitzung (eine Unterrichtsstunde)
Vortestung	Experiment	Follow-up
Erfassung relevanter Hintergrund-/ Kontrollvariablen: 1. Lesegeschwindigkeit, 2. figurale Analogiebildung, 3. Leseinteresse, 4. Selbstwirksamkeit, 5. Zielorientierungen	1. Erfassung der situationsspezifischen Testangst in Antizipation des Experiments, 2. Experimentelle Variation der Treatmentbedingungen, über **5 Texte** mit insgesamt **37 Items**, 3. Erfassung der Testangst nach dem Experiment, 4. Erfassung der wahrgenommenen Nützlichkeit der Feedbacks (nur für Experimentalgruppen)	Erfassung der Lesekompetenz: **4 Texte** mit insgesamt **30 Items** (neues, aber zu Experiment und Posttest vergleichbares Material)
	Posttest	
	5. Erfassung der Lesekompetenz: **2 Texte** mit insgesamt **14 Items** (neues, aber zum Experiment vergleichbares Material)	

Durchschnittlich eine Woche nach der Vortestung fand die zweite Sitzung, das computerbasierte Experiment mit anschließendem Posttest, statt. Diese Sitzung wurde in den Computerräumen der teilnehmenden Schulen durchgeführt. In einigen Schulen

kamen zusätzlich Laptops zum Einsatz, wenn in einer Schule nicht ausreichend Computer vorhanden waren.
Die Schüler wurden den fünf Versuchsbedingungen randomisiert zugewiesen. Der Testleiter gab anfangs eine kurze Einleitung zum Zweck der Untersuchungssitzung. Die eigentliche Instruktion für die Bearbeitung von Fragebogen, Experiment und Posttest erfolgte über das Computerprogramm selbst. Die Entscheidung, die Instruktion computerbasiert zu geben, war mit Rücksicht auf die Kontrollgruppe, die kein Treatment erhielt, getroffen worden. Sie sollte nicht durch die Erläuterungen für die Experimentalgruppen bezüglich der Feedbackgabe beeinflusst werden.

Der Testleiter war während der gesamten Sitzung anwesend. Für die Sitzung stand eine doppelte Unterrichtsstunde zur Verfügung und es wurde ohne Unterbrechung gearbeitet. Die Dauer der offiziellen Schulpause durfte normalerweise zur Untersuchungszeit dazugeschlagen werden. Somit standen in der Regel 90 Minuten plus die Pausenzeit von etwa 10 Minuten zur Verfügung. Wenn Probanden am Ende der zweiten Schulstunde den Posttest noch nicht beendet hatten, konnten sie auf eigenen Wunsch noch einige Minuten in der anschließenden Pause weiterarbeiten. Als Testleiter fungierten zum größten Teil studentische Hilfskräfte, die im Vorfeld der Untersuchung geschult wurden. Sämtliche Instruktionen, die durch den Testleiter gegeben wurden, waren dem Manual nach wörtlich vorzutragen, wodurch die Standardisierung der einzelnen Untersuchungsdurchführungen gewahrt wurde.

Durchschnittlich vier Wochen nach dem Experiment fand das Follow-up statt. Diese Sitzung markierte das Ende der gesamten Untersuchung. Für das Follow-up wurde aus organisatorischen Gründen auf ein Papier-und-Bleistift-basiertes Vorgehen zurückgegriffen. Die Untersuchungsdurchführung mit Testheft konnte in den normalen Klassenräumen der Schulen umgesetzt werden und das erleichterte die Organisation des Follow-ups sehr. Die Sitzung verlief ähnlich wie die Vortestung. Am Anfang erläuterte ein Testleiter den Zweck der Sitzung und gab die Instruktion zur Bearbeitung des Tests. Um gegenseitiges Abschreiben zu unterbinden, erhielten nebeneinandersitzende Schüler abwechselnd Form A und B des Testhefts. Das Follow-up wurde in einer Unterrichtsstunde durchgeführt.

Alle Untersuchungssitzungen fanden während der regulären Unterrichtszeit statt. Am Ende jeder Sitzung erhielten die Teilnehmer eine kleine Aufmerksamkeit (z.B. eine Süßigkeit, einen Stift) als Dankeschön für ihre Teilnahme. Die Untersuchungen wurden hauptsächlich von studentischen Hilfskräften durchgeführt, die im Vorfeld der Untersuchung eine ausführliche Schulung durchliefen.

5.8 Datenanalyse

5.8.1 Fehlende Werte und Fallausschluss

Für alle drei Untersuchungstermine ergaben sich Stichprobenausfälle, die hauptsächlich auf krankheitsbedingte Abwesenheiten der Schüler zurückgingen. Die Stichprobenumfänge der einzelnen Untersuchungssitzungen sind in Tabelle 15 zusammengefasst. Von zentraler Bedeutung ist die Gruppe der N = 566 Probanden, die am Experiment und am Posttest teilgenommen haben. Für einen kleinen Teil dieser Teilstichprobe liegen keine Daten aus der Vortestung und/oder dem Follow-up vor.

Da die Instrumente der Vortestung und des Follow-ups in Papier-und-Bleistift-Form administriert wurden, können sich hier fehlende Werte dadurch ergeben, dass Items nicht oder nicht eindeutig beantwortet wurden. Darüber hinaus bleiben drei der insgesamt N = 609 Fälle der Vortestung (0.5 %) aufgrund instruktionswidrigen Verhaltens für die Auswertungen unberücksichtigt. Für das Follow-up gibt es keine entsprechenden Ausschlüsse.

Tabelle 15 Stichprobenumfänge zu den einzelnen Untersuchungssitzungen

Vortest	Experiment und Posttest	Follow-up	n
✓	✓	✓	505
✓	✓	-	27
-	✓	✓	23
-	✓	-	11
✓	-	✓	60
✓	-	-	17
-	-	✓	21
n = 609	n = 566	n = 609	N = 664

Anmerkungen. ✓ = teilgenommen; - = nicht teilgenommen.

In den computeradministrierten Lesekompetenztests der Treatmentphase und des Posttests sowie den dabei ebenfalls computerbasiert eingesetzten Fragebogen treten fehlende Werte dann auf, wenn das Programm nicht weiter bearbeitet wurde. Von den insgesamt N = 566 Probanden, die an der zweiten Sitzung teilnahmen, liegt für N = 560 Fälle (98.94 %) ein vollständiger Datensatz vor. Damit steht von insgesamt sechs Probanden ein unvollständiger Datensatz zur Verfügung, wovon sich in fünf Fällen die Ausfälle bereits im Experiment ergaben und in einem weiteren Fall der Posttest begonnen, aber nicht fertig bearbeitet wurde. Die unvollständigen Datensätze verteilen sich annähernd gleich auf die experimentellen Bedingungen (Kein Feedback: n = 2; Knowledge of Result: n = 1; Fehlererklärung: n = 2; Inferenzprompt: n = 1 und metakognitiver Prompt: n = 0). Die fehlenden Werte werden nicht als Fehler gewertet, aber auch nicht ersetzt, sondern die betreffenden Fälle werden aus den jeweiligen Analysen des Experiments bzw. des Posttests ausgeschlossen.

Des Weiteren werden Fälle ausgeschlossen, in denen zu viele Lesekompetenzitems „durchgeklickt“, das heißt schneller als erwartet beantwortet wurden, da hier eine hinreichend ernsthafte Testbearbeitung in Frage zu stellen ist. Ausgehend von Erfahrungswerten durch Ausprobieren wird davon ausgegangen, dass für die meisten der eingesetzten Lesekompetenzitems bei guter Lesegeschwindigkeit mindestens sechs Sekunden gebraucht werden, um sie jeweils annähernd vollständig lesen und eine Antwortauswahl treffen zu können. Das bedeutet ausdrücklich nicht, dass diese Zeit für die Mehrheit der Leser als ausreichend betrachtet wird, um ein Item zu lesen und durch Nachdenken oder Nachlesen im Text eine begründete Auswahl aus den Antwortalternativen zu treffen. Die Festsetzung von sechs Sekunden ist als Minimalkriterium zu verstehen. Folglich gelten Aufgaben, die in weniger als sechs Sekunden beantwortet wurden, als auffällig schnell beantwortet und wahrscheinlich nicht hinreichend bearbeitet. Für jeden Untersuchungsteilnehmer ist berechnet, wie viele Items des Experiments und des Posttests in unter sechs Sekunden beantwortet wurden. Die entsprechenden Häufigkeitsverteilungen sind in Tabelle 16 aufgeführt.
Die Entscheidung für den Cut-off-Wert von sechs Sekunden beruht zwar auf Erfahrungswerten, dennoch ist dieses Kriterium zu diskutieren. Deshalb wurden die Daten auch unter drei alternativen, höheren Cut-off-Werten analysiert. Nach diesen alternativen Kriterien gelten Items als zu schnell beantwortet, wenn sie in weniger als sieben, acht bzw. neun Sekunden beantwortet wurden. Diese Kriterien sind im Vergleich

zu dem Kriterium von weniger als sechs Sekunden strenger, weil sie die Zeitspanne, bis wann ein Item als zu schnell beantwortet gilt, höher ansetzen. Die entsprechenden Häufigkeitsverteilungen für die drei Kriterien sind ebenfalls in Tabelle 16 aufgeführt. Diese Gegenüberstellung der Kriterien verschafft einen Überblick über die Daten und zeigt, dass die höheren Cut- off-Werte zu keiner deutlichen Anhebung der Fälle führen, in denen viele Items „durchgeklickt" wurden. Außerdem zeigen die Häufigkeitsverteilungen, dass die (große) Mehrheit der Teilnehmer immer länger für die Beantwortung von Items aufgebracht hat als sechs, sieben, acht bzw. neun Sekunden (vgl. Tabelle 16, erste Zeile), das heißt kein Item „durchgeklickt" wurde – egal welches der Cut-off-Kriterien angewendet wurde. Einzelne durchgeklickte Items kommen eher vor als viele durchgeklickte Items pro Person.

Für den Fallausschluss wurde letztlich also der Cut-off-Wert von sechs Sekunden (d.h. Antwort in weniger als 6 Sekunden gilt als „durchgeklickt") beibehalten und keiner der alternativen, höheren Werte[6]. Ausgeschlossen wurden nun die Fälle, in denen *mehrere* Items in weniger als sechs Sekunden beantwortet wurden. Ein gewisser Spielraum „durchgeklickter" Items pro Person erscheint hier sinnvoll, denn unter gewissen Umständen ist eine sehr schnelle Antwort nachvollziehbar. Es wird festgelegt, dass im Experiment (N = 37 Items) bis zu fünf „durchgeklickte" Items erlaubt sein sollen. Fünf von 37 Items entsprechen einem prozentualen Anteil von annähernd 14 % des Tests. Dieser Spielraum für wahrscheinlich zu schnell beantwortete Items erscheint vertretbar und wird höheren Quoten (6/37 = 16.2 %; 7/37 = 18.9 %; 8/37 = 21.6 %; 9/37 = 24.3 %; 10/37 = 27.0 %; 11/37 = 29.7 %) vorgezogen. Für den Posttest bedeutet das, dass hier maximal zwei der insgesamt 14 Items (= 14.3 %) „durchgeklickt" worden sein dürfen.
Von den insgesamt N = 566 Probanden des Experiments sind für die Analysen der Treatmentphase demzufolge N = 66 Fälle auszuschließen. Diese Fälle entstammen zu etwa gleichen Teilen allen fünf Versuchsgruppen (χ^2 = .82; p > .05) und allen drei Schulformen (χ^2 = 4.36; p > .05). Der entsprechende Fallausschluss aus dem Posttest umfasst N = 187 Fälle. Diese verteilen sich ebenfalls gleich auf die fünf Versuchsgruppen (χ^2 = 2.33; p > .05), aber ungleich auf die drei Schulformen (χ^2 = 16.82; p < .001). Dabei

[6] Die zentralen Analysen dieses Experiments wurden jedoch zusätzlich unter der Nutzung der drei alternativen Cut-off-Werte durchgeführt. Die Ergebnisse dieser Analysen (siehe Anhang B) widersprechen nicht den im Hauptteil der Arbeit berichteten Ergebnissen.

zeigt sich, dass fast doppelt so viele Hauptschüler (N = 88) wie Realschüler (N = 55) oder Gymnasiasten (N = 44) im Posttest vermehrt schneller bzw. zu schnell ihre Antworten auf die Fragen abgegeben haben.

Tabelle 16 Häufigkeit „durchgeklickter“ Items

		Cut-off-Werte, *unter denen* ein Item als „zu schnell beantwortet“ gilt									
	Menge der Items[a]	< 6 Sekunden (= gewähltes Kriterium)			< 7 Sekunden			< 8 Sekunden			< 9 Sekunden		
		N	(%)		N	(%)		N	%		N	%	
Test der Treatmentphase (37 Items)	0	402	(71,0)	Σ =500	373	(65,9)	Σ =490	351	(62,0)	Σ = 476	316	(55,8)	Σ = 466
	1	33	(5,8)		46	(8,1)		52	(9,2)		66	(11,7)	
	2	14	(2,5)		25	(4,4)		27	(4,8)		32	(5,7)	
	3	24	(4,2)		18	(3,2)		24	(4,2)		28	(4,9)	
	4	16	(2,8)		13	(2,3)		13	(2,3)		16	(2,8)	
	5	11	(1,9)		15	(2,7)		9	(1,6)		8	(1,4)	
	6	11	(1,9)	Σ = 66	11	(81,9)	Σ =76	14	(2,5)	Σ =90	13	(2,3)	Σ=100
	7	4	(0,7)		7	(1,2)		13	(2,3)		13	(2,3)	
	8	8	(1,4)		8	(1,4)		9	(1,6)		11	(1,9)	
	9	4	(0,7)		6	(1,1)		4	(0,7)		7	(1,2)	
	10	3	(0,5)		5	(0,9)		6	(1,1)		6	(1,1)	
	>10	21	(3,8)		25	(4,5)		28	(5,1)		33	(5,8)	
	>20	15	(2,8)		14	(2,6)		16	(3,0)		17	(3,2)	
	>30	0	(0,0)		0	(0,0)		0	(0,0)		0	(0,0)	
	Σ	566	(100)		566	(100)		566	(100)		566	(100)	
Posttest (14 Items)	0	313	(55,3)	Σ =379	286	(50,5)	Σ =363	256	(45,2)	Σ =345	213	(37,6)	Σ =324
	1	43	(7,6)		52	(9,2)		59	(10,4)		75	(13,3)	
	2	23	(4,1)		25	(4,4)		30	(5,3)		36	(6,4)	
	3	16	(2,8)	Σ =187	14	(2,5)	Σ =203	21	(3,7)	Σ =221	27	(4,8)	Σ =242
	4	16	(2,8)		18	(3,2)		17	(3,0)		14	(2,5)	
	5	15	(2,7)		13	(2,3)		13	(2,3)		22	(3,9)	
	6	16	(2,8)		13	(2,3)		20	(3,5)		18	(3,2)	
	7	9	(1,6)		14	(2,5)		10	(1,8)		12	(2,1)	
	8	21	(3,7)		17	(3,0)		8	(1,4)		9	(1,6)	
	9	17	(3,0)		20	(3,5)		20	(3,5)		17	(3,0)	
	10	20	(3,5)		19	(3,4)		21	(3,7)		21	(3,7)	
	>10	57	(10,0)		75	(13,2)		91	(16,2)		102	(18,0)	
	Σ	566	(100)		566	(100)		566	(100)		566	(100)	

Anmerkungen. [a] In der Spalte sind die Anzahl der Items, auf die das jeweilige Cut-off-Kriterium zutrifft, abgetragen (d.h. wie oft wurden Items unterhalb der zeitlichen Grenzwerte beantwortet).

Der Umfang des Fallausschlusses für die Leistungsanalysen des Posttests ist erheblich; er macht mit den N = 187 Fällen immerhin etwa ein Drittel der Gesamtstichprobe aus. Auch wenn die Höhe des Fallausschlusses und mögliche Gründe für die vermehrten kürzeren Antwortzeiten zu diskutieren sein werden, wird dennoch an den Ausschlusskriterien festgehalten. Der gewählte Cut-off-Wert von unter sechs Sekunden wird als eine zeitliche Mindestanforderung für das Lesen und Beantworten einer durchschnittlichen Aufgabe verstanden – schnellere, vor allem viele schnellere Antworten können selbst bei guten und sehr guten Lesefähigkeiten nicht Ausdruck einer begründeten Antwort sein. Eine hinreichend ernsthafte Testbearbeitung stellt aber die Grundlage für die Wirkung der Interventionen in diesem Experiment dar. Die Alternative zum Ausschluss der vielen Fälle im Posttest würde darin bestehen, alle Fälle zu berücksichtigen oder die Höhe der Fallausschlüsse zu reduzieren, indem ein niedrigerer Cut-off-Wert, unter dem ein Item als zu schnell beantwortet gilt, genutzt und/oder die Menge an zugelassenen „durchgeklickten" Items erhöht werden würde. Beide Alternativen stehen aber konträr zu den zuvor gegebenen Begründungen *für* den Fallausschluss anhand der letztlich genutzten Kriterien. Somit wird an dem Fallausschluss von N = 187 Fällen im Posttest festgehalten.

Auswertungsstichproben: Für die Analysen der Leistungen im Textverständnis/der Lesekompetenz werden letztlich also die Fälle einbezogen, für die aus der Treatmentphase bzw. dem Posttest ein vollständiger Datensatz vorliegt und die mit dem gewählten Cut- off-Wert nicht als „Durchklicker" klassifiziert sind. Damit ergibt sich für die Treatmentphase eine Auswertungsstichprobe von N = 495. Für den Posttest, für den nicht nur die „Durchklicker" des Posttests, sondern logischerweise auch die des Experiments auszuschließen sind, resultiert eine Auswertungsstichprobe von N = 365.
Für die Auswertung des Follow-ups werden die Fälle ausgeschlossen, für die Hinweise vorliegen, dass das Treatment nicht bzw. nicht ausreichend durchlaufen wurde. Damit sind zum einen die Fälle der „Durchklicker" aus der Experimentalphase auszuschließen und zum anderen die Fälle unvollständiger Datensätze, weil diese zum Teil erheblich weniger Test- und Lerngelegenheiten ausgesetzt waren als Probanden, die die gesamte Intervention bearbeitet haben. Daraus resultiert für das Follow-up eine Auswertungsstichprobe von N = 524.

5.8.2 Scoring

Die Antworten in den Lesekompetenztests des Experiments, des Posttests und des Follow-ups wurden stets als richtig/falsch mit der Codierung 1/0 gewertet. Im Experiment wurden die ersten und zweiten Antwortversuche der Feedbackbedingungen dabei gleichwertig behandelt – eine richtige Antwort im zweiten Versuch wurde ebenfalls mit 1 gewertet. Die Einzelantworten wurden zu Summenscores aggregiert. Somit ergibt sich hier für alle experimentellen Bedingungen ein Summenwert für die ersten Versuche und für die Feedbackbedingungen zusätzlich ein Summenwert für die zweiten Versuche. Der Summenwert der zweiten Versuche steht allerdings immer in Abhängigkeit vom Summenwert der ersten Versuche, denn aus der Menge der Fehler in den Erstantworten ergibt sich die Anzahl benötigter zweiter Versuche. Weiterhin können alle Versuchsgruppen in den Summenscores für den Posttest und das Follow-up verglichen werden.

5.8.3 Auswertungsplan

Die im Anschluss dargestellten Analysen und Ergebnisse richten sich im Wesentlichen nach dem folgenden Auswertungsplan: In einem ersten Schritt wurde die „Qualität“ der in der Untersuchung eingesetzten Lesekompetenztests untersucht. Das Material ist die Grundlage der Intervention und der elaborierten Feedbacks, die zum Teil auf die Fragestellungen bzw. Antwortalternativen ausgerichtet wurden. Damit sollten die Items, die letztlich die Grundlage für die Auswertungen und Interpretationen bilden, zumindest gewissen Minimalkriterien entsprechen. Die Lesekompetenzitems wurden dazu mittels Item- bzw. Skalenanalysen im Rahmen der Klassischen Testtheorie sowie Raschanalysen zur Überprüfung der Dimensionalität analysiert und gänzlich unpassende Items wurden aus den weiteren Analysen ausgeschlossen. Die Untersuchung verfolgte nicht das Ziel einer Testentwicklung und dementsprechend wurden eher liberalere Richtlinien zur Beurteilung der Item- bzw. Modellkennwerte angelegt.

Für die Item- bzw. Skalenanalysen wurden nur die Daten der Kontrollgruppe (N = 75), nicht die der Feedbackbedingungen herangezogen. Denn die Feedbackbedingungen waren per se einer Intervention ausgesetzt, die die Testleistung positiv und/oder negativ (und zwischen den Feedbackgruppen möglicherweise auch in unterschiedlicher Richtung) beeinflusst hat. Um aber eine „Baseline“ der Leistungsfähigkeit der gewählten Schüler-

bzw. Altersgruppe bestimmen zu können, brauchte es eine standardisierte Testbedingung, wie sie die feedbackfreie Kontrollgruppe bietet.

Im zweiten Schritt wurden die Versuchsgruppen hinsichtlich der zur Verfügung stehenden Person- bzw. Hintergrundvariablen auf Gleichverteilung getestet. Diese Überprüfung ergänzt die vorgenommene Randomisierung, die in den jeweilig untersuchten Gruppen/Klassenverbänden eines Untersuchungstermins durchgeführt wurde.

Danach wurden die Fragestellungen dieses Experiments bezüglich der Wirksamkeit der Feedbackinterventionen untersucht. Dabei wurden nacheinander zunächst die Feedbackeffekte auf die Leistung in der Treatmentphase, dem Posttest und dem Follow-up analysiert. Im Anschluss daran wurden die Bearbeitungszeiten und mögliche Effekte der Feedbackinterventionen darauf untersucht. Schließlich wurden noch mögliche Auswirkungen des Treatments auf die Testangst der Probanden analysiert. Zur Auswertung der Feedbackeffekte kamen hauptsächlich varianzanalytische Verfahren zum Einsatz, die mit der Software SPSS ausgeführt wurden.

6 Ergebnisse

6.1 Beschreibung der Lesekompetenzitems

Zunächst wurden alle Lesekompetenzitems der Treatmentphase und des Posttests (insgesamt N = 51 Items) mittels Item- und Reliabilitätsanalyse im Rahmen der Klassischen Testtheorie analysiert. Die Grundlage der Berechnung stellt die Kontrollgruppe dar, deren Umfang nach Ausschließen von „Durchklickern“ und unvollständigen Datensätzen (vgl. Abschnitt 5.8.1) N = 75 beträgt. Die Analysen zeigen, dass fast alle Items (N = 49 Items) im mittleren Schwierigkeitsbereich von $p = .20$ bis $p = .80$ (Bühner, 2006) liegen. Zwei Items weisen eine extreme Schwierigkeit von $p = .16$ (Item mit Label Kli2) und $p = .17$ (Kak7) auf. Die mittlere Itemschwierigkeit beträgt $p = .46$. Die interne Konsistenz ist mit Cronbachs $\alpha = .84$ (Präzision $P_\alpha < .01$) als gut zu bewerten. Die mittlere Inter-Item-Korrelation (MIC) wiederum fällt mit einem Wert von .09 gering aus.

Hinsichtlich ihrer Trennschärfen waren die Items auffällig: zwei der 51 Items wiesen negative Trennschärfen auf ($r = -.15$ für Kli2 und $r = -.09$ für Spr6) und für 14 weitere Items ergeben sich mit $r < .19$ sehr niedrige Zusammenhänge (Ebel, 1979) zu den restlichen Lesekompetenzitems. Die inhaltliche Analyse dieser Items zeigt in Abgrenzung zu den anderen keine spezifischen Abweichungen in den Fragestellungen bzw. der Art der gestellten Anforderung. Tendenziell scheint es sich hier um Items zu handeln, die komplexere Sachverhalte abfragen, für die Kinder der sechsten Klassenstufe möglicherweise nicht zuverlässig eine passende Vorstellung aufbauen können, weil ihnen entsprechende Erfahrungen fehlen.

Die Items mit niedrigen, aber positiven Trennschärfen werden letztlich im Itempool belassen, auf dessen Grundlage die Analysen der Effekte der Feedbackinterventionen durchgeführt werden. Die zwei Items mit negativen Trennschärfen (Kli2 und Spr6), die aus dem Testteil der Treatmentphase stammen, werden dagegen ausgeschlossen. Auch wenn sie inhaltlich zu den formulierten Anforderungen passen, stehen sie dennoch in einem negativen Zusammenhang zu den restlichen Items und würden daher die zu bildenden Leistungssummenscores eher negativ beeinflussen.

Die verbleibenden 49 Items aus der Treatmentphase (nunmehr 35 Items) und dem Posttest (14 Items) wurden einer erneuten Reliabilitätsanalyse unterzogen. Die mittlere Itemschwierigkeit liegt weiterhin bei $p = .46$ und damit im mittleren Schwierigkeitsbereich. Es verbleiben 14 Items mit sehr niedrigen Trennschärfen von $r < .19$. Die Inter-Item-Korrelation (MIC) beträgt .11 und ist damit leicht angestiegen. Das Cronbachs Alpha steigt durch den Ausschluss der zwei Items leicht an auf $\alpha = .85$. Mit einer Präzision von $P_\alpha < .01$ spricht das Alpha für eine gute interne Konsistenz der verbliebenen Lesekompetenzitems.

Die verbliebenen 49 Items aus Treatment und Posttest wurden zusätzlich im Rahmen der Raschskalierung mittels Conquest (Wu, Adams, Wilson & Haldane, 2007) analysiert. Als Richtwerte zur Beurteilung auffälliger Kennwerte werden ein MNSQ kleiner 0.80 oder größer 1.20 sowie ein *T*-Wert kleiner - 2.0 und größer 2.0 herangezogen (Adams, 2002). Zwei der 49 Items haben einen MNSQ von 0.80 und einen signifikant abweichen *T*-Wert (Item Nat6 mit $T = - 2.7$ und MNSQ = 0.80; Item Man3 mit $T = - 2.5$ und MNSQ = 0.80). Die Kennwerte sprechen in beiden Fällen für keine gute Passung zum Modell. Deshalb werden beide Items, eines aus dem Test des Treatments und eines aus dem Posttest, im nächsten Schritt ebenfalls ausgeschlossen. Alle anderen Items liegen bei mindestens einem der Kennwerte innerhalb des akzeptablen Bereichs.
Die verbliebenden N = 47 Items aus Treatmentphase (nunmehr 34 Items) und Posttest (nunmehr 13 Items) wurden erneut raschskaliert. Diese Auswertung bringt kein Item mit auffällig abweichenden MNSQ- und *T*-Werten hervor. Es werden folglich keine weiteren Items ausgeschlossen. Die Item- und Modellkennwerte der N = 47 Items aus Treatment und Posttest sind in Tabelle 17 aufgeführt. Die Tabelle enthält ebenfalls die Kennwerte für Itemschwierigkeit und Itemtrennschärfe aus der Klassischen Testtheorie. Darüber hinaus ist in Abbildung 4 die latente Verteilung der Personen- und der Itemparameter auf einer gemeinsamen Logit-Skala dargestellt.
Die Itemmenge von N = 47 Items ist die Grundlage für die in den späteren Abschnitten dargelegten Analysen der Feedbackeffekte. Die Skalenkennwerte ergeben sich durch den Ausschluss der vier Items wie folgt: die interne Konsistenz ist mit Cronbachs $\alpha = .84$ (Präzision $P_\alpha < .01$) unverändert und als gut zu bewerten; die mittlere Itemschwierigkeit beträgt ebenfalls weiterhin $p = .46$ und die mittlere Inter-Item-Korrelation ist mit MIC = .10 minimal erhöht.

Tabelle 17 Itemkennwerte

Nr	Label	δ	MNSQ	T	p	r
Experiment						
1	Ant1	0.72	1.11	1.0	.31	.11
2	Ant2	-0.76	1.03	0.3	.63	.28
3	Ant3	0.86	0.99	-0.1	.28	.29
4	Ant4	0.72	1.11	1.0	.31	.09
5	Ant5	-0.04	0.90	-1.4	.47	.44
6	Ant6	0.21	1.15	1.9	.41	.06
7	Kli1	-0.83	0.93	-0.8	.64	.42
8	Kli3	-0.39	1.09	1.2	.55	.13
9	Kli4	-0.27	1.01	0.2	.52	.30
10	Kli5	0.59	0.94	-0.6	.33	.36
11	Kli6	0.39	1.15	1.7	.37	.07
12	Spr1	0.94	0.99	-0.1	.27	.26
13	Spr2	-0.76	0.95	-0.5	.63	.39
14	Spr3	-0.70	1.15	1.6	.61	.07
15	Spr4	0.46	1.11	1.2	.36	.11
16	Spr5	-0.70	0.91	-1.1	.61	.45
17	Spr7	1.01	1.08	0.6	.25	.14
18	Spr8	-1.31	0.84	-1.2	.73	.52
19	Som1	-0.89	0.83	-1.7	.65	.51
20	Som2	-0.04	1.09	1.1	.47	.16
21	Som3	0.08	0.95	-0.6	.44	.38
22	Som4	-0.28	1.05	0.7	.52	.21
23	Som5	-0.64	0.90	-1.1	.60	.44
24	Som6	1.01	1.06	0.5	.25	.13
25	Som7	0.46	1.10	1.1	.36	.13
26	Som8	-0.52	0.97	-0.3	.57	.32
27	Som9	0.08	1.00	-0.0	.44	.28
28	Nat1	0.72	1.02	0.2	.31	.24
29	Nat2	0.33	0.96	-0.5	.39	.38
30	Nat3	-0.34	1.00	0.1	.53	.30
31	Nat4	0.46	0.99	-0.1	.36	.28
32	Nat5	-0.83	0.94	-0.6	.64	.38
33	Nat7	-0.96	0.82	-1.9	.67	.60
34	Nat8	0.59	1.08	0.8	.33	.14
Posttest						
35	Man1	-0.89	0.84	-1.7	.65	.55
36	Man2	0.72	1.09	0.9	.31	.11
37	Man4	-0.21	0.98	-0.2	.51	.29
38	Man5	-0.33	0.91	-1.3	.53	.43
39	Man6	-0.76	1.02	0.2	.63	.25
40	Man7	-0.70	0.87	-1.5	.61	.48
Noch Posttest						
41	Kak1	-0.33	0.87	-1.8	.53	.49
42	Kak2	-0.09	0.97	-0.4	.48	.35
43	Kak3	0.52	1.00	0.0	.35	.29
44	Kak4	0.27	1.03	0.4	.40	.19
45	Kak5	0.27	1.05	0.6	.40	.23
46	Kak6	0.65	1.13	1.2	.32	.07
47	Kak7	1.53	0.95	-0.2	.17	.33
Follow-up						
1	Ele2	-0.00	0.94	-0.6	.64	.45
2	Ele3	0.14	1.04	0.4	.61	.32
3	Ele4	0.75	1.09	1.1	.49	.25
4	Ele5	0.65	0.92	-1.0	.50	.47
5	Ele6	0.14	1.17	1.7	.61	.21
6	Tan1	0.88	1.06	0.7	.47	.36
7	Tan2	0.16	1.08	0.9	.60	.31
8	Tan3	-0.62	0.86	-1.1	.78	.51
9	Tan5	1.79	1.15	1.4	.33	.19
10	Tan6	1.32	0.91	-1.0	.38	.44
11	Tan7	0.83	1.07	0.8	.48	.31
12	Fan1	0.33	1.01	0.2	.58	.36
13	Fan2	-1.14	0.98	-0.1	.84	.31
14	Fan3	0.51	1.05	0.6	.54	.31
15	Fan4	-2.73	1.03	0.2	.96	.12
16	Fan5	-0.69	1.03	0.3	.78	.28
17	Fan6	-1.25	1.00	0.0	.87	.38
18	Fan7	-1.16	0.86	-0.8	.85	.47
19	Fan8	-0.45	0.95	-0.4	.72	.43
20	Fan9	-0.04	1.06	0.6	.66	.39
21	Bah1	0.75	1.09	1.0	.52	.25
22	Bah2	-0.73	0.96	-0.3	.77	.44
23	Bah3	-0.09	0.99	-0.0	.66	.40
24	Bah4	0.94	1.00	0.0	.46	.41
25	Bah5	-0.28	0.95	-0.4	.72	.41
26	Bah6	0.17	0.86	-1.6	.67	.56
27	Bah7	0.17	1.04	0.5	.47	.30
28	Bah8	0.91	0.84	-1.2	.80	.50

Anmerkungen. Label = gibt die Unit (vgl. Tabelle 13) und die Nummer des Items darin an; δ = Itemschwierigkeit (Raschmodell); MNSQ = Weighted Mean Square (Raschmodell); T = Wert aus T-Verteilung (Raschmodell); p = Itemschwierigkeit (klassische Testtheorie); r = Itemtrennschärfe (klassische Testtheorie).

Die Lesekompetenzitems des Follow-ups sind separat von den Lesekompetenzitems des Experiments und Posttests ausgewertet. Die Datengrundlage ist die Leistung der

Probanden, die zuvor im Experiment in der Kontrollbedingung gearbeitet haben (N = 105). Die Item- und Reliabilitätsanalysen im Sinne der Klassischen Testtheorie ergibt, dass fünf der insgesamt 30 Items des Follow-ups mit jeweils $p > .80$ eine extreme Schwierigkeit aufweisen. Die restlichen Items liegen mit $.20 < p > .80$ im mittleren Schwierigkeitsbereich. Die mittlere Itemschwierigkeit beträgt $p = .62$. Ferner weisen drei Items einen Trennschärfekoeffizienten von $r < .19$ und damit einen sehr niedrigen Zusammenhang (Ebel, 1979) zu den restlichen Items des Follow-ups auf. Die mittlere Inter-Item-Korrelation (MIC) ist mit .15 entsprechend niedrig. Die Reliabilität ist, gemessen als interne Konsistenz, mit einem Cronbachs $\alpha = .84$ (Präzision $P_{\alpha} < .01$) als gut zu bewerten.

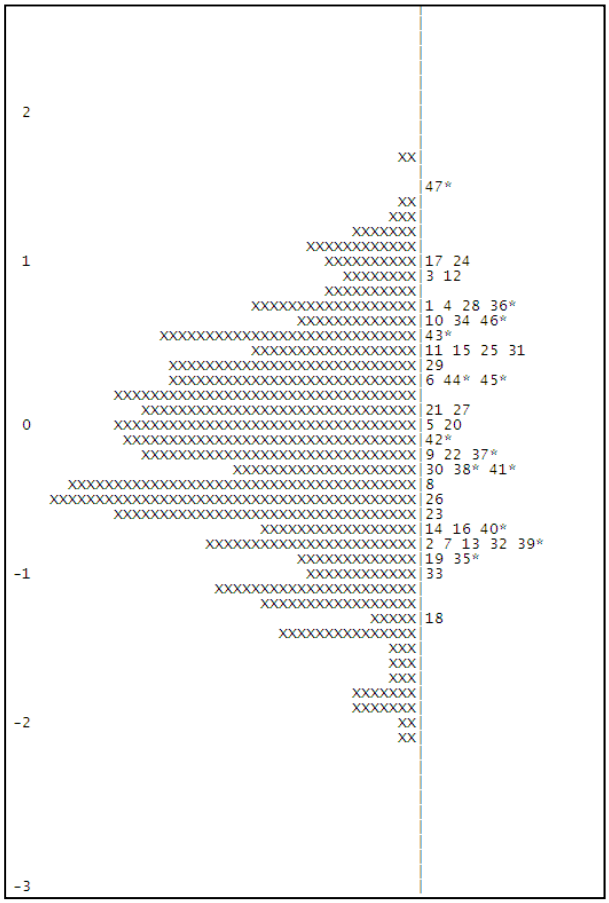

Abbildung 4 Lesekompetenzitems aus Treatment und Posttest: Latente Verteilung der Personenparameter (Kreuze links) und Itemparameter (Zahlen rechts) auf einer gemeinsamen Logit-Skala.

Anmerkungen. Jedes 'x' repräsentiert 0.1 Fälle. Die mit * gekennzeichneten Items gehören zum Posttest; die hier abgebildeten Itemnummern korrespondieren mit der Itemnummerierung in Tabelle 17.

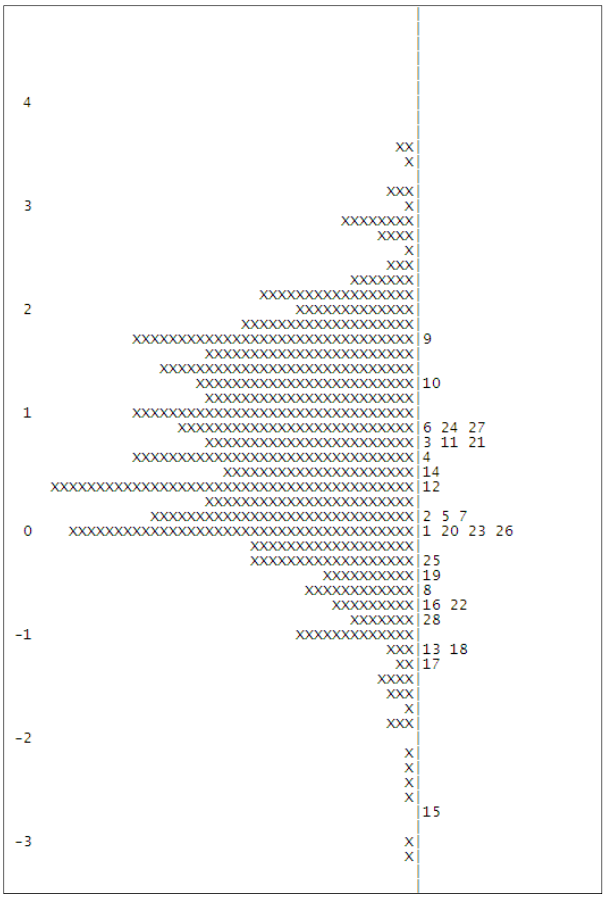

Abbildung 5 Lesekompetenzitems des Follow-ups: Latente Verteilung der Personenparameter (Kreuze links) und Itemparameter (Zahlen rechts) auf einer gemeinsamen Logit-Skala.

Anmerkungen. Jedes 'x' repräsentiert 0.1 Fälle. Die hier abgebildeten Itemnummern korrespondieren mit der Itemnummerierung in Tabelle 17/Tabellenabschnitt Follow-up.

Die Raschskalierung der N = 30 Items des Follow-ups zeigt wiederum, dass zwei Items auffällig abweichende Kennwerte von 1.20 < MNSQ < 0.80 und 2.0 < *T* < - 2.0 aufweisen (Item Ele1 mit Item Tan4). Nach Ausschluss dieser Items und erneuter Skalierung der verbliebenen N = 28 Items zeigen sich keine weiteren Auffälligkeiten. Die Kennwerte dieser Raschskalierung sind ebenfalls in Tabelle 17 aufgeführt und ihnen sind wiederum die Kennwerte der Itemschwierigkeit und der Itemtrennschärfe aus der Reliabilitätsanalyse nach der Klassischen Testtheorie zur Seite gestellt. Abbildung 5 illustriert die auf der Raschskalierung der N = 28 Items basierende latente Verteilung der Personen- und der Itemparameter auf einer gemeinsamen Logit-Skala dargestellt.

Durch den Ausschluss der zwei Items haben sich die Skalenkennwerte des Follow-up-Tests minimal geändert: die interne Konsistenz weist weiterhin ein Cronbachs α = .84

($P_\alpha < .01$) auf, die mittlere Itemschwierigkeit ist minimal erhöht auf $m = .63$, ebenso die mittlere Inter-Item-Korrelation MIC = .16. Darüber hinaus ist ein Item mit einer Trennschärfe $r < .19$ verblieben. Im Vergleich zum Itemmaterial des Experiments mit anschließendem Posttest (mittlere Itemschwierigkeit $m = .46$) weist der Test des Follow-ups mit der mittleren Itemschwierigkeit von $m = .63$ eine geringere Schwierigkeit auf.

Die Ergebnisse der Item- und Reliabilitätsanalysen und der Raschskalierungen zusammengenommen widersprechen nicht der Annahme, dass es sich bei den Lesekompetenzitems des Experiments/Posttests bzw. des Follow-ups um ein eindimensionales Konstrukt handelt. Die Leistungen in Experiment, Posttest und Follow-up werden zu jeweils einem Summenscore zusammengefasst.

6.2 Überprüfung von a-priori Gruppenunterschieden

Die Versuchsgruppen wurden, ergänzend zur vorgenommenen Randomisierung der Untersuchungsteilnehmer, im Vorfeld der Leistungsanalysen auf zufällig aufgetretene Gruppenunterschiede in den erhobenen leistungsrelevanten Hintergrundvariablen untersucht. Für die Prüfung auf Gruppenhomogenität standen sowohl die Leistungstests und die motivational-emotionalen Fragebogen als auch die Personmerkmale Geschlecht und Sprache sowie die Schulform zur Verfügung. Alle diese Merkmale können einen Einfluss auf die Lesekompetenz und/oder die Feedbackwirksamkeit ausüben.
Die Leistungstests und die Fragebogenmaße wurden jeweils mittels ANOVA analysiert. Es zeigen sich keine statistisch signifikanten Unterschiede zwischen den Versuchsgruppen, weder in der Lesegeschwindigkeit ($F\,(4, 459) = 0.28$; $p > .05$; $\eta^2 = .002$) noch in der figuralen Intelligenz ($F\,(4, 459) = 0.90$; $p > .05$; $\eta^2 = .008$). Auch in den motivational-emotionalen Maßen, das heißt dem Leseinteresse ($F\,(4, 459) = 0.34$; $p > .05$; $\eta^2 = .003$), der Selbstwirksamkeit im Bereich des Lesens ($F\,(4, 449) = 0.96$; $p > .05$; $\eta^2 = .008$), der Aufgabenorientierung ($F\,(4, 459) = 1.02$; $p > .05$; $\eta^2 = .009$) und der Ichorientierung ($F\,(4, 459) = 0.67$; $p > .05$; $\eta^2 = .006$) sowie der Testangst hinsichtlich Besorgtheit ($F\,(4, 490) = 0.11$; $p > .05$; $\eta^2 = .001$) und Aufgeregtheit ($F\,(4, 490) = 0.33$; $p > .05$; $\eta^2 = .003$), zeigen sich a-priori keine signifikanten Gruppenunterschiede. Die Mittelwerte und Standardabweichungen der Versuchsgruppen in den berücksichtigten Hintergrundvariablen sind in Tabelle 18 zusammengestellt.

Tabelle 18 Deskriptive Statistiken der erfassten Hintergrundvariablen

		N	*M*	*SD*
Lesegeschwindigkeit	Kein Feedback	94	35.49	6.88
(N = 464)	Knowledge of Result	90	35.63	8.20
	Metakognitiver Prompt	96	34.84	7.63
	Fehlererklärung	92	34.93	6.91
	Inferenzprompt	92	35.73	7.46
Figurale Intelligenz	Kein Feedback	94	15.70	6.27
(N = 464)	Knowledge of Result	90	16.62	6.07
	Metakognitiver Prompt	96	16.20	6.20
	Fehlererklärung	92	17.27	6.57
	Inferenzprompt	92	16.84	5.54
Leseinteresse	Kein Feedback	94	2.81	0.79
(N = 464)	Knowledge of Result	90	2.86	0.76
	Metakognitiver Prompt	96	2.73	0.86
	Fehlererklärung	92	2.78	0.75
	Inferenzprompt	92	2.82	0.79
Selbstwirksamkeit	Kein Feedback	88	3.04	0.46
(N = 454)	Knowledge of Result	90	3.09	0.48
	Metakognitiver Prompt	94	3.08	0.43
	Fehlererklärung	91	3.01	0.45
	Inferenzprompt	91	2.98	0.46
Zielorientierungen/	Kein Feedback	94	3.18	0.59
Aufgabenorientierung	Knowledge of Result	90	3.22	0.55
(N = 464)	Metakognitiver Prompt	96	3.13	0.61
	Fehlererklärung	92	3.19	0.54
	Inferenzprompt	92	3.07	0.49
Zielorientierungen/	Kein Feedback	94	2.70	0.82
Ichorientierung	Knowledge of Result	90	2.78	0.72
(N =464)	Metakognitiver Prompt	96	2.82	0.78
	Fehlererklärung	92	2.65	0.86
	Inferenzprompt	92	2.72	0.80
Testangst/Besorgtheit	Kein Feedback	95	2.29	0.67
(erfasst vor dem Experiment)	Knowledge of Result	96	2.30	0.67
(N = 495)	Metakognitiver Prompt	109	2.32	0.70
	Fehlererklärung	96	2.34	0.70
	Inferenzprompt	99	2.34	0.68
Testangst/Aufgeregtheit	Kein Feedback	95	1.61	0.62
(erfasst vor dem Experiment)	Knowledge of Result	96	1.68	0.71
(N = 495)	Metakognitiver Prompt	109	1.65	0.56
	Fehlererklärung	96	1.69	0.68
	Inferenzprompt	99	1.62	0.56

Die Ausprägungen in den Personmerkmalen Geschlecht und Sprache sowie die Schulformen werden mittels Chiquadrat-Tests auf Gleichverteilung zwischen den

Versuchsgruppen getestet. Die Ergebnisse belegen, dass sowohl der Mädchen- und Jungenanteil (χ^2 = 1.85, p > .05), deutsche und nicht deutsche Herkunftssprache (χ^2 = 2.17, p > .05) als auch die drei Schulformen (χ^2 = 5.75, p > .05) zwischen den Versuchsgruppen gleichverteilt sind. Die absoluten Häufigkeiten der Merkmalsausprägungen der drei Variablen sind in Tabelle 19 dargestellt.

Tabelle 19 Häufigkeiten für Personmerkmale und Schulformen in den Versuchsgruppen

	Knowledge of Result	Metakogni-tiver Prompt	Fehler-erklärung	Inferenz-prompt	Kein Feedback	Σ
Geschlecht						
Mädchen	52	56	57	56	49	270
Jungen	44	53	39	43	46	225
Σ	96	109	96	99	95	495
Sprache						
Deutsch	93	101	89	89	89	26
andere	3	8	5	6	4	461
Σ	96	109	94	95	93	495
Schulformen						
Hauptschule	33	40	35	36	34	178
Realschule	27	42	28	35	30	162
Gymnasium	36	27	33	28	31	155
Σ	96	109	96	99	95	495

A-priori besteht zwischen den Versuchsgruppen also keine Ungleichverteilung in einer der für die Lesekompetenz und/oder die Feedbackwirksamkeit relevanten, in dieser Untersuchung zur Verfügung stehenden Hintergrund-/Personvariablen. Folglich sollten potentielle Feedbackeffekte auf die Treatmentbedingungen zurückführbar sein.

6.3 Haupteffekte von Feedback auf die Leistung

Im Folgenden werden die Analysen zur zentralen Fragestellung nach der Wirksamkeit der verschiedenen Feedbackbedingungen auf das Textverständnis/die Lesekompetenz berichtet. Die Auswirkungen auf die Leistung wurden bezüglich der Erst- und Zweitantworten aus der Treatmentphase sowie des unmittelbaren Posttests und des Follow-ups untersucht.

6.3.1 Die Leistung in der Treatmentphase

Erstantworten

Die Verwendung der Leistungen in den Erstantworten im Lesekompetenztest der Treatmentphase resultierte aus der Annahme, dass sich ein möglicher Nutzen von Feedback bereits im Verlauf der Interventionsphase, das heißt für nachfolgende Items, die ähnliche Anforderungen an den Leser stellen, zeigt. Die varianzanalytische Auswertung der Erstantworten zeigt, dass kein Haupteffekt des Feedbackfaktors auf die Leistung in den Erstantworten nachweisbar ist (F (4, 490) = 0.54; $p > .05$; $\eta^2 = .004$). Die mittleren Leistungen aller Bedingungen fallen, wie in Tabelle 20 ersichtlich, sehr ähnlich aus (vgl. auch Abbildung 7).

Die Kennwerte der deskriptiven Statistik basieren auf der Anzahl richtig gelöster Aufgaben im ersten Versuch. Somit weist die deskriptive Statistik in Tabelle 20 implizit ebenso daraufhin, dass bei den Erstantworten im Durchschnitt etwa 19 bis 20 Fehler gemacht wurden. Das entspricht einer durchschnittlichen Fehlerquote von annähernd 55 % bis 58 %. Für die Feedbackbedingungen bedeutet das, dass die Probanden bei 34 Items im Durchschnitt etwa 20 Feedbackgaben erhielten.

Tabelle 20 Deskriptive Statistik der Erstantworten in der Treatmentphase (N = 495)

		Leistung in den Erstantworten (34 Items)				
	N	*M*	*SD*	*M* %	*Min*	*Max*
Kein Feedback	95	15.36	5.46	45.18	3	26
Knowledge of Result	96	14.38	5.09	42.29	5	25
Metakognitiver Prompt	109	14.43	5.49	42.44	4	27
Fehlererklärung	96	14.52	5.35	42.71	4	28
Inferenzprompt	99	14.61	5.48	42.97	2	30

Nach der Analyse der Erstantworten anhand des Summenscores wurde die Leistung der Versuchsgruppen in den Erstantworten zusätzlich über den Verlauf des Treatments untersucht. Es sollte ausgeschlossen werden, dass durch eventuell aufgetretene Leistungsschwankungen in den Feedbackinterventionen (z.B. anfangs schlechtere, gegen Ende des Treatments bessere Leistung als die Kontrollgruppe) und die Verwendung des Gesamtsummenscores mögliche positive Effekte in Teilen des Treatments verdeckt blieben. Als sozusagen „natürliche“ Analyseeinheiten boten sich hier die einzelnen fünf

Units der Treatmentphase an. Um die Vergleichbarkeit zwischen den Units zu erhöhen, wurden nicht die absoluten Werte pro Unit, sondern die an der Anzahl der Items pro Unit relativierten Lösungshäufigkeiten verglichen. Da die Units des Experiments in zwei Reihenfolgeversionen vorgegeben worden waren, wurden die Analysen getrennt für diese durchgeführt.

Die Analyse mittels MANOVAs ergab keine signifikanten Unterschiede zwischen den Versuchsgruppen (für die jeweiligen F-Statistiken vgl. Tabelle 21). Die deskriptiven Statistiken sind in Tabelle 21 aufgeführt und die Mittelwerte der relativen Lösungshäufigkeiten sind zusätzlich in Abbildung 6 illustriert. Wie in der Abbildung ersichtlich, ist eine Unterscheidung der durchschnittlichen Leistungen der Versuchsgruppen auch über die fünf Units hinweg nicht möglich.

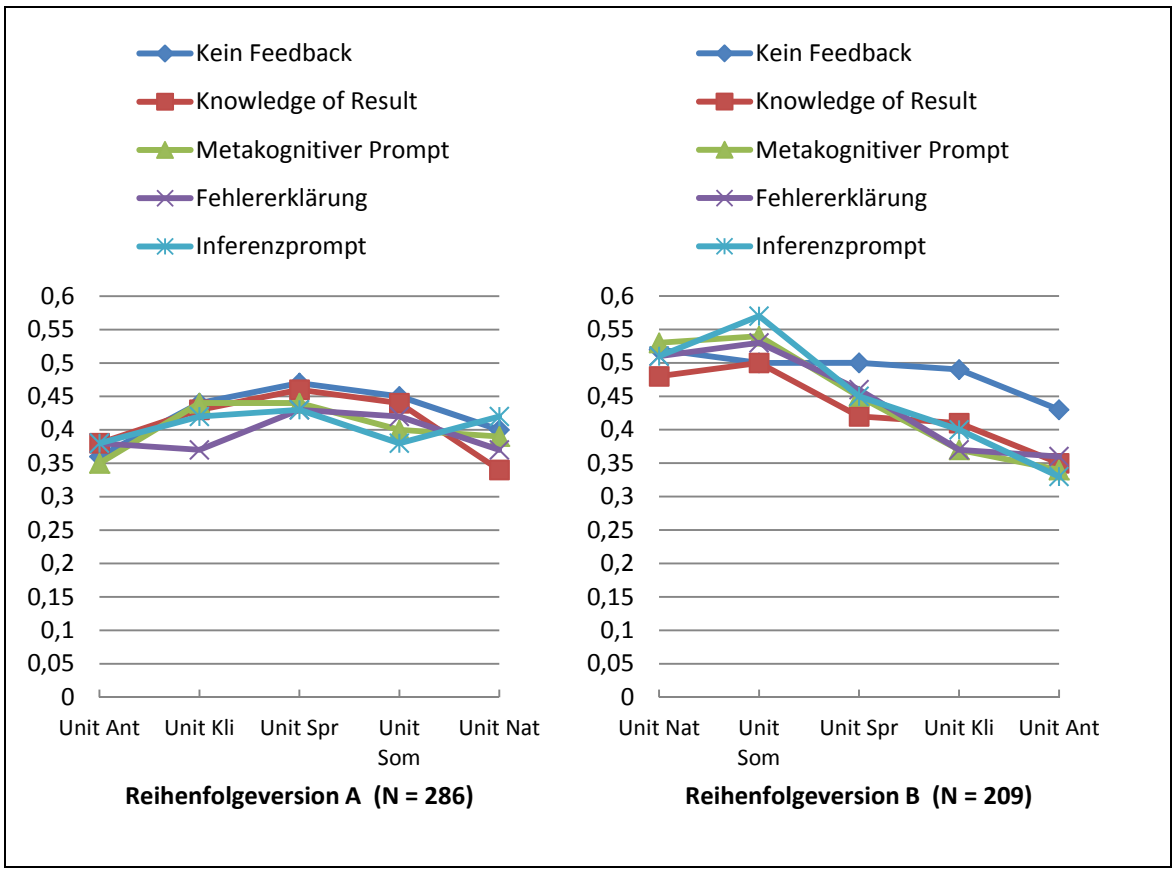

Abbildung 6 Relative Lösungshäufigkeiten in den Erstantworten pro Unit, getrennt für beide Reihenfolgeversionen des Experiments.

Tabelle 21 MANOVAs für Erstantworten pro Unit (N = 495)

Unit	Versuchsgruppen	Reihenfolgeversion A			Reihenfolgeversion B		
		N	*M*	*SD*	N	*M*	*SD*
Antarktis	Kein Feedback	59	0.36	0.22	36	0.43	0.20
(1./5. Position)	Knowledge of Result	55	0.38	0.21	41	0.35	0.24
	Metakognitiver Prompt	68	0.35	0.21	41	0.34	0.19
	Fehlererklärung	49	0.38	0.21	47	0.36	0.25
	Inferenzprompt	55	0.38	0.21	44	0.33	0.24
		F (4, 281) = 0.23; $p > .05$ $\eta^2 = .00$			F (4, 204) = 1.16; $p > .05$ $\eta^2 = .02$		
Klimawandel	Kein Feedback	59	0.44	0.24	36	0.49	0.26
(2./4. Position)	Knowledge of Result	55	0.43	0.28	41	0.41	0.29
	Metakognitiver Prompt	68	0.44	0.25	41	0.37	0.25
	Fehlererklärung	49	0.37	0.31	47	0.37	0.26
	Inferenzprompt	55	0.42	0.29	44	0.40	0.27
		F (4, 281) = 0.65; $p > .05$ $\eta^2 = .01$			F (4, 204) = 1.28; $p > .05$ $\eta^2 = .02$		
Sprung	Kein Feedback	59	0.47	0.23	36	0.50	0.24
(3./3. Position)	Knowledge of Result	55	0.46	0.20	41	0.42	0.23
	Metakognitiver Prompt	68	0.44	0.21	41	0.45	0.21
	Fehlererklärung	49	0.43	0.21	47	0.46	0.22
	Inferenzprompt	55	0.43	0.21	44	0.45	0.24
		F (4, 281) = 0.43; $p > .05$ $\eta^2 = .01$			F (4, 204) = 0.66; $p > .05$ $\eta^2 = .01$		
Sommer	Kein Feedback	59	0.45	0.25	36	0.50	0.23
(4./2. Position)	Knowledge of Result	55	0.44	0.20	41	0.50	0.22
	Metakognitiver Prompt	68	0.40	0.22	41	0.54	0.26
	Fehlererklärung	49	0.42	0.23	47	0.53	0.20
	Inferenzprompt	55	0.38	0.22	44	0.57	0.21
		F (4, 281) = 1.01; $p > .05$ $\eta^2 = .01$			F (4, 204) = 0.63; $p > .05$ $\eta^2 = .01$		
Nationalpark	Kein Feedback	59	0.40	0.24	36	0.52	0.19
(5./1. Position)	Knowledge of Result	55	0.34	0.20	41	0.48	0.23
	Metakognitiver Prompt	68	0.39	0.26	41	0.53	0.26
	Fehlererklärung	49	0.37	0.22	47	0.51	0.24
	Inferenzprompt	55	0.42	0.22	44	0.51	0.22
		F (4, 281) = 1.03; $p > .05$ $\eta^2 = .01$			F (4, 204) = 0.22; $p > .05$ $\eta^2 = .00$		

Anmerkungen. Die Werte basieren auf den richtigen Antworten im Erstversuch, die zum Zweck der Vergleichbarkeit jeweils an der Anzahl der Items der Unit relativiert wurden.

Die Analyseergebnisse bezüglich der Erstantworten in der Treatmentphase belegen damit, dass sich die Gabe der drei elaborierten Feedbackarten nicht zu einer Leistungssteigerung in den Erstversuchen in der Treatmentphase führte. Dieses Ergebnis widerspricht den a-

priori formulierten Annahmen. Erwartungskonform ist dagegen, dass sich das einfache Feedback Knowledge of Result und die feedbackfreie Kontrollgruppe hinsichtlich der Leistung in den Erstantworten nicht voneinander unterscheiden.

Zweitantworten

Für die Feedbackbedingungen liegen aus der Treatmentphase neben den Antworten in den Erstversuchen auch jene der Zweitversuche vor. Diese informieren darüber, in welchem Ausmaß zunächst falsch beantwortete Aufgaben nach den entsprechenden Feedbackmitteilungen richtig gelöst wurden. Dabei ist die Feedbackbedingung Knowledge of Result die Bedingung, der gegenüber der mögliche Mehrwert elaborierter Feedbackarten zu bewerten ist.

Die deskriptive Statistik der Zweitantworten ist in Tabelle 22 aufgeführt (vgl. auch Abbildung 7). Zur Beurteilung möglicher Gruppenunterschiede ist in erster Linie das Maß der an der Anzahl der benötigten zweiten Versuche relativierten Lösungshäufigkeiten heranzuziehen (schattierte Spalten in Tabelle 22). Diesbezüglich wird bereits aus dem numerischen Vergleich der deskriptiven Kennwerte deutlich, dass sich die Feedbackgruppen nicht oder nur geringfügig voneinander abheben. Die ANOVA belegt entsprechend, dass es keinen statistisch signifikanten Haupteffekt des Feedbacks auf die Leistung in den Zweitversuchen gibt (F (3, 396) = 0.86; $p > .05$; $\eta^2 = .006$).

Tabelle 22 Deskriptive Statistik der Zweitantworten in der Treatmentphase (N = 400)

		Leistung in den Zweitantworten						
		Absolute Häufigkeiten		Relative Häufigkeiten [a]				
	N	*M*	*SD*	*M*	*SD*	*M* %	*Min*	*Max*
Knowledge of Result	96	8.38	3.00	0.45	0.16	44.49	0.09	0.77
Metakognitiver Prompt	109	8.08	2.72	0.44	0.16	43.76	0.13	0.86
Fehlererklärung	96	7.78	2.54	0.42	0.14	41.71	0.17	0.77
Inferenzprompt	99	7.79	2.67	0.42	0.14	41.82	0.09	0.82

Anmerkungen. [a] Die Summe richtiger Antworten im 2. Versuch ist relativiert an der Anzahl der benötigten 2. Versuche.

Die Ergebnisse bezüglich der Zweitantworten widerlegen damit die a-priori formulierten Annahmen, da keine der elaborierten Feedbackarten zu einer höheren Korrekturleistung

führte als sie im Zusammenhang mit dem Feedback Knowledge of Result zu beobachten ist.

In Abbildung 7 sind die durchschnittlichen Leistungen der experimentellen Bedingungen in den Erstversuchen und, die Kontrollgruppe ohne Feedback ausgeschlossen, in den Zweitversuchen. Diese gemeinsame Darstellung der durchschnittlichen Gruppenleistung in beiden Maße illustriert das Ausmaß, in dem durch das Einräumen der zweiten Antwortmöglichkeit nach Feedback die Gesamttestleistung verbessert wurde.

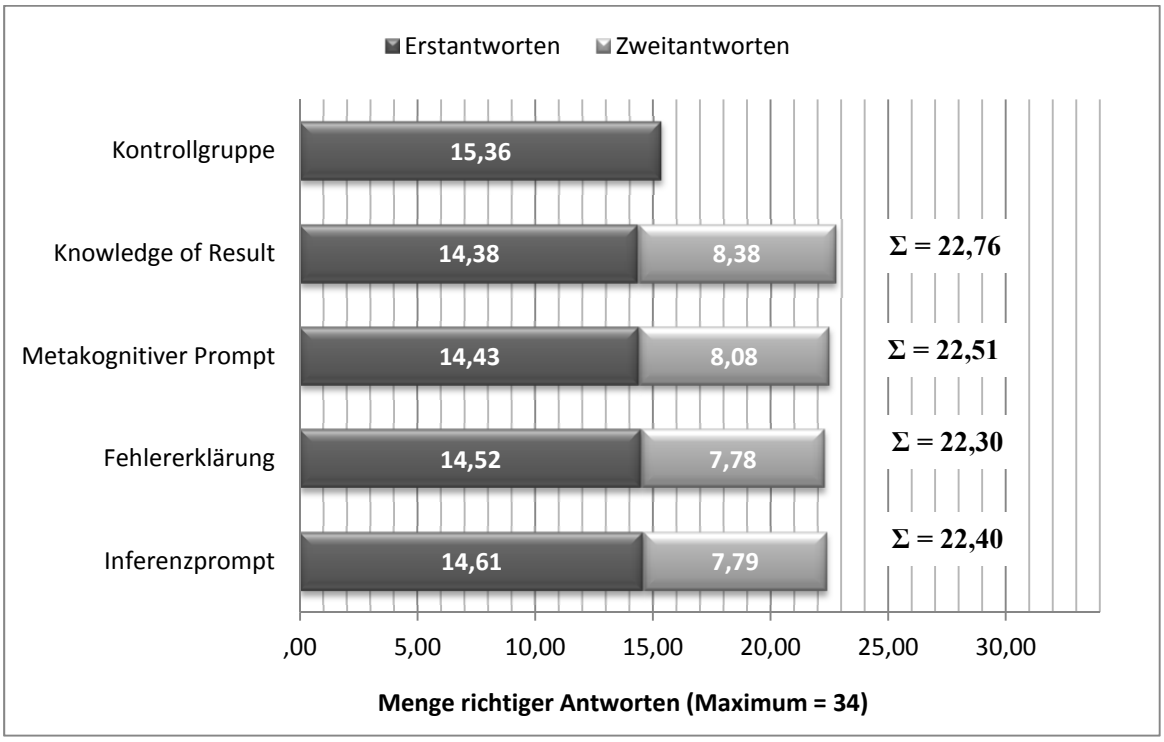

Abbildung 7 Gemeinsame Darstellung der mittleren Leistungen in den Erst- und Zweitantworten der Treatmentphase.

6.3.2 Die Leistung im Posttest

Nach den Leistungen in der Treatmentphase ist als abhängige Variable die Leistung im unmittelbaren Posttest zu überprüfen. Wie ein Blick auf die deskriptive Statistik zeigt (vgl. Tabelle 23) fallen die mittleren Leistungen der Versuchsgruppen sehr ähnlich aus. Es wurden im Durchschnitt jeweils annähernd fünf bis sechs der insgesamt 13 verwendeten Posttestitems richtig gelöst. Die ANOVA belegt, dass es keinen statistisch signifikanten Haupteffekt des Feedbacks auf die Posttestleistung gibt (F (4, 360) = 0.83; $p > .05$; $\eta^2 = .009$).

Tabelle 23 Deskriptive Statistik für die Leistung im Posttest (N = 365)

	Leistung im Posttest (13 Items)					
	N	*M*	*SD*	*M* %	*Min*	*Max*
Kein Feedback	75	5.89	2.72	45.31	1	13
Knowledge of Result	71	6.01	2.70	46.23	1	11
Metakognitiver Prompt	81	5.75	2.60	44.23	2	12
Fehlererklärung	62	5.66	2.95	43.54	0	12
Inferenzprompt	76	5.25	2.93	40.39	1	13

Damit sind die a-priori formulierten Annahmen über die Wirkung der elaborierten Feedbackarten auch für die Leistung im Posttest widerlegt. Keine der elaborierten Feedbackarten führt zu einer Leistungssteigerung. Erwartungskonform zeigt sich dagegen, dass Knowledge of Result keine Transferleistungen nach sich zieht.

6.3.3 Vergleich der Leistungen in der Treatment- und der Posttestphase

Zur Analyse möglicher Unterschiede in den Leistungen in beiden Testteilen der experimentellen Sitzung, das heißt der Treatmentphase (Erstantworten) und dem Posttest, sowie potentieller Abhängigkeiten von den Versuchsgruppen wurde eine ANOVA mit Messwiederholung (RM-ANOVA) mit den Versuchsgruppen als Zwischensubjektfaktor und den Messzeitpunkten (Treatmentphase vs. Posttest) als Innersubjektfaktor durchgeführt. Die Leistungen gingen aufgrund der unterschiedlichen Anzahl an Items in der Treatmentphase und dem Posttest als relative Lösungshäufigkeiten in die Analysen ein.

Die RM-ANOVA ergibt einen signifikanten Haupteffekt des Innersubjektfaktors auf die Leistung (F (1, 360) = 4.80; $p < .05$; $\eta^2 = .01$) – die Posttestleistung fällt im Durchschnitt geringer aus als die Leistung in den Erstversuchen der Treatmentphase. Signifikante Unterschiede zwischen den Versuchsgruppen (Zwischensubjektfaktor) sind nicht nachweisbar (F (4, 360) = 0.36; $p > .05$; $\eta^2 = .004$), ebenso kein signifikanter Interaktionseffekt aus Messzeitpunkt und Versuchsgruppen (F (4, 360) = 1.58; $p > .05$; partielles $\eta^2 = .02$). Die Mittelwerte und Standardabweichungen der relativen Lösungshäufigkeiten sind in Tabelle 24 zusammengefasst und Abbildung 8 veranschaulicht die mittleren Ausprägungen der Gruppen zu beiden Messzeitpunkten.

Tabelle 24 Deskriptive Statistik der relativen Lösungshäufigkeiten im Treatment und im Posttest sowie der Differenz daraus (N = 365)

		Treatment (Erst-antworten)		Posttest		Differenzwert (Treatment - Posttest)			
	N	*M*	*SD*	*M*	*SD*	*M*	*SD*	*T*	*p*
Kein Feedback	75	0.47	0.17	0.45	0.21	0.01	0.16	0.66	>.05
Knowledge of Result	71	0.45	0.15	0.46	0.21	-0.01	0.16	-0.45	>.05
Metakognitiver Prompt	81	0.45	0.17	0.44	0.20	0.01	0.16	0.35	>.05
Fehlererklärung	62	0.47	0.16	0.44	0.23	0.03	0.18	1.40	>.05
Inferenzprompt	76	0.46	0.15	0.40	0.23	0.05	0.17	2.77	<.01

Tabelle 24 enthält zudem als Ergänzung zur RM-ANOVA die Gruppenmittelwerte für die Differenzwerte der Leistungen in beiden Testteilen und die entsprechenden T-Werte für die Mittelwertsvergleiche beider Testteile pro Versuchsgruppe. Ein positiver Differenzwert bedeutet, dass die relative Leistung im Posttest geringer ausfällt als in der Treatmentphase; ein negativer Differenzwert steht entsprechend für das gegenteilige Ereignis. Anhand der mittleren Differenzwerte ist ersichtlich, dass außer der Bedingung Knowledge of Result alle Gruppen im Durchschnitt im Posttest verhältnismäßig weniger Aufgaben richtig gelöst haben als in der Treatmentphase (positive Differenzwerte). Die Abweichungen sind jedoch statistisch nicht signifikant, ausgenommen für die Bedingung Inferenzprompt. Für diese Gruppe ergibt sich auf der Grundlage der Differenzwerte eine signifikant niedrigere Leistung im Posttest als in den Erstversuchen der Treatmentphase. Auf diesen Unterschied geht im Wesentlichen auch der signifikante Haupteffekt des Innersubjektfaktors (Messzeitpunkte) der RM-ANOVA zurück. Der höhere Differenzwert der Gruppe Inferenzprompt drückt sich im Vergleich zu den Kontrollbedingungen (Cohens d = 0.24 zu Kontrollgruppe; d = 0.36 zu Knowledge of Result) in kleinen Effektstärken (d > 0.20) nach Cohen (1992) aus.

6.3.4 Die Leistung im Follow-up

Für die Bewertung der Feedbackeffekte auf die Lesekompetenz bleibt als letzte abhängige Variable die Leistung im Follow-up zu prüfen. Die deskriptive Statistik hierzu ist in Tabelle 25 zusammengefasst. Die Mittelwerte der Versuchsgruppen fallen wie in den zuvor berichteten abhängigen Variablen sehr ähnlich aus. Die ANOVA belegt, dass kein Haupteffekt des Feedbackfaktors auf die Leistung im Follow-up vorliegt (F (4, 519) = 0.31; p > .05; η^2 = .002). Damit bringt keine der elaborierten Feedbackbedingungen

durchschnittlich vier Wochen nach dem Experiment (noch) einen Effekt hervor und die a-priori getroffenen Annahmen zu den Auswirkungen der elaborierten Feedbackarten sind auch hier widerlegt. Die Annahme bezüglich des Knowledge of Result trifft wiederum zu.

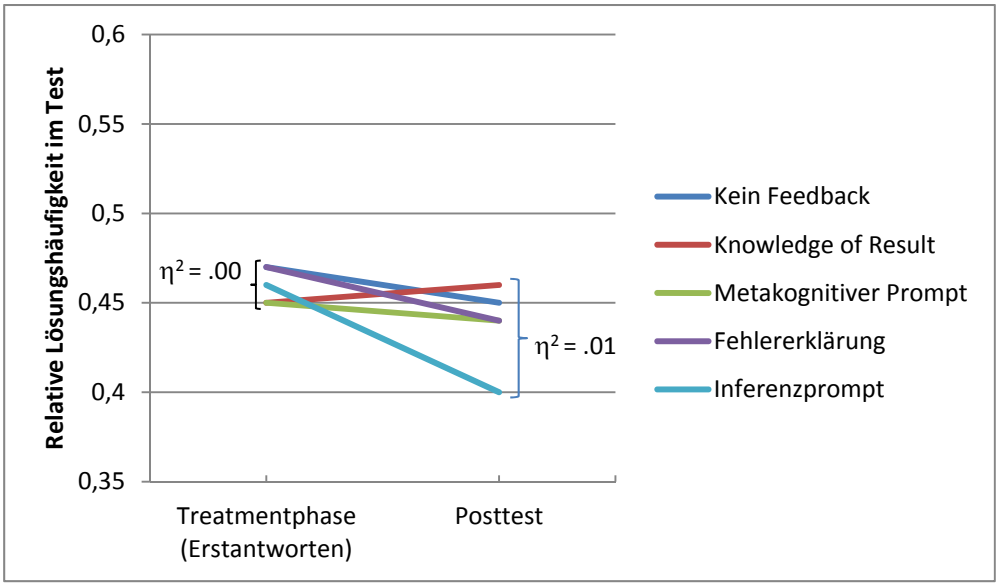

Abbildung 8 Durchschnittliche relative Lösungshäufigkeiten des Tests der Treatmentphase (Erstantworten) und des Posttests.

Tabelle 25 Deskriptive Statistik für die Leistung im Follow-up (N = 524)

	Leistung im Follow-up (28 Items)					
	N	*M*	*SD*	*M* %	*Min*	*Max*
Kein Feedback	104	17.45	5.68	62.32	1	27
Knowledge of Result	106	17.16	5.32	61.29	6	27
Metakognitiver Prompt	104	17.18	5.48	61.36	4	28
Fehlererklärung	103	16.67	5.77	59.54	2	27
Inferenzprompt	107	16.87	5.38	60.25	4	28

6.3.5 Zusammenfassung der Ergebnisse zu den Feedbackeffekten auf das Textverständnis bzw. die Lesekompetenz

Zusammenfassend ist für die Analysen der Lesekompetenztests festzuhalten, dass keine der (elaborierten) Feedbackinterventionen einen statistisch signifikanten Effekt auf die Leistung in den Erst- oder Zweitantworten in der Treatmentphase oder dem Posttest oder dem Follow-up nach sich gezogen hat. Die Anzahl der „Treatmentmomente“

(Feedbackgaben), die sich in dieser Untersuchung aus der Anzahl der Fehler bei den Erstantworten ergibt, beläuft sich im Durchschnitt auf etwa 20 Rückmeldungen.

6.4 Haupteffekt von Feedback auf die Bearbeitungszeiten

Die Auswertung der deskriptiven Statistik der Bearbeitungszeiten hat zunächst ergeben, dass sowohl für das Experiment als auch den Posttest Ausreißerwerte vorliegen. Als Ausreißerwerte sind hier Werte definiert, die das Anderthalbfache des Interquartilsabstandes (IQA) unterhalb des ersten Quartils (Q1) oder oberhalb des dritten Quartils (Q3) liegen ($Q3 + 1.5*IQA < x < Q1 - 1.5*IQA$). Wie in Abbildung 9 und Abbildung 10 ersichtlich, treten hier nur Ausreißerwerte oberhalb des dritten Quartils, das heißt Bearbeitungszeiten, die in Relation zu den anderen Werten extrem lang ausfallen, auf. Bezogen auf das Experiment trifft dies auf fünf Probanden zu. Sie haben zwischen 85.65 Minuten und 97.09 Minuten für diesen Testteil aufgewendet. Beim Posttest resultieren die Ausreißerwerte aus drei Fällen, die zwischen 64.36 Minuten und 71.99 Minuten für den Posttest benötigt haben.
Die Inspektion dieser Einzelfälle von Extremwerten hat ergeben, dass sie im Posttest ausnahmslos durch extrem lange Verweilzeiten bei einem einzelnen, und zwar dem letzten Item zustande kommen. Dieser Umstand ist mit einer hohen Wahrscheinlichkeit darauf zurückzuführen, dass diese Probanden den Posttest zwar bereits fertig bearbeitet, aber die letzte erforderliche Aktion, das Speichern der Ergebnisse, durch die das Programm und die Zeiterfassung abgeschlossen wurde, noch nicht bzw. nicht selbstständig ausgeführt hatten, sondern erst nach Aufforderung durch den Testleiter.

In Bezug auf die Treatmentphase suggerieren die Daten, dass die Ausreißerwerte hier ebenfalls durch außergewöhnlich hohe Verweilzeiten bei einzelnen Items, typischerweise jene am Ende der Treatmentphase, begründet sind. Die Ursache hierfür kann aber weniger „technischer“ Art wie im Posttest sein. Stattdessen liegt als Erklärung nah, dass die Probanden pausiert hatten, denn die Bearbeitungszeiten der nachfolgenden Items sind dann unauffällig. Aber auch wenn in einzelnen Fällen das Zustandekommen der Ausreißerwerte nachvollziehbar ist und das dafür verantwortliche Test- bzw. Arbeitsverhalten nicht gegen die Zuverlässigkeit der Daten dieser Probanden spricht,

führen die Extremwerte dennoch zu Verzerrungen in den Analysen der Bearbeitungszeiten und werden dafür ausgeschlossen.

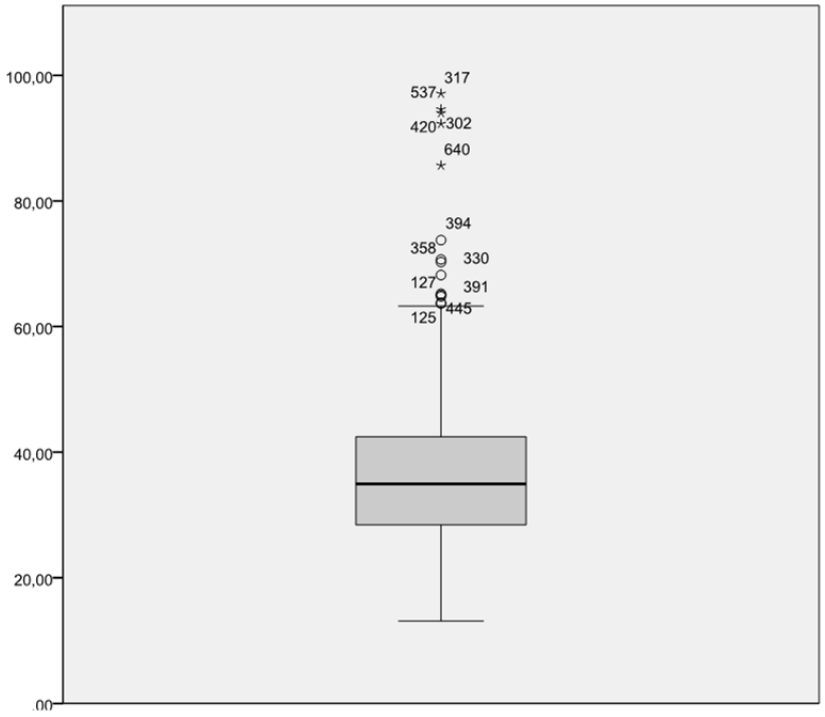

Abbildung 9 Box-Plot der Bearbeitungszeit für das Experiment (N = 495).

Anmerkungen. * = Ausreißerwert; ° = Extremwert.

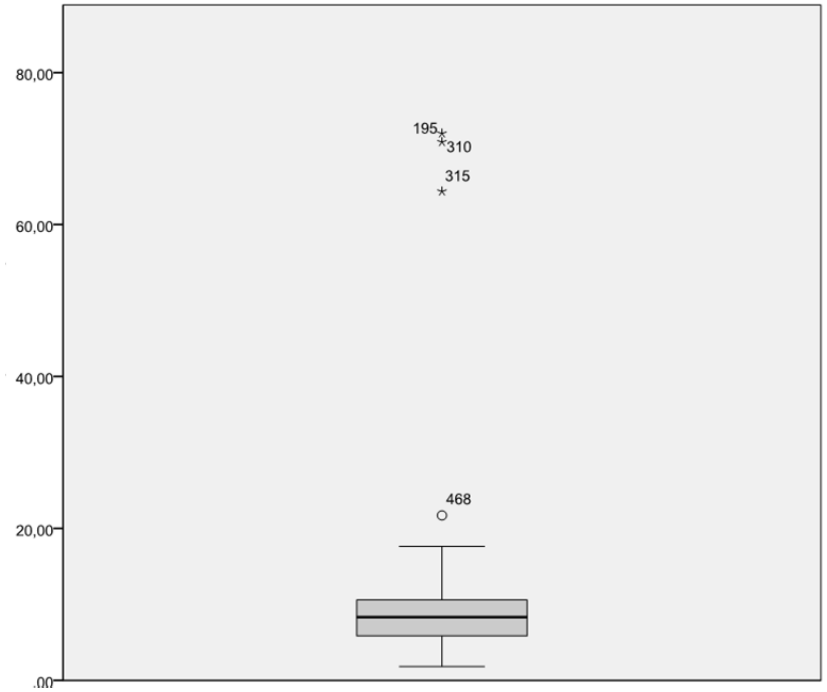

Abbildung 10 Box-Plot der Bearbeitungszeit für den Posttest (N = 365).

Anmerkungen. * = Ausreißerwert; ° = Extremwert.

Treatmentphase

Zur Beantwortung der zweiten Fragestellung nach möglichen Auswirkungen der Feedbackinterventionen auf die Bearbeitungsdauer der Items wurden die Versuchsgruppen zunächst hinsichtlich der Zeit, die sie für das gesamte Experiment aufgewendet haben, verglichen. Hier zeigt die deskriptive Statistik (vgl. Tabelle 26), dass von den Feedbackbedingungen im Durchschnitt zumindest numerisch mehr Zeit für das *gesamte* Experiment aufgewendet wurde als von der Kontrollgruppe ohne Feedback. Die ANOVA zeigt aber, dass kein statistische signifikanter Haupteffekt des Feedbacks auf die Gesamtbearbeitungszeit der Treatmentphase vorliegt (F (4, 485) = 1.07; $p > .05$; $\eta^2 = .01$).

Im nächsten Schritt wurde die Gesamtbearbeitungszeit in die Minuten, die auf alle Erstversuche entfallen, und die Zeit, die in den Feedbackbedingungen für die Zweitantworten genutzt wurde, aufgeteilt und analysiert. Diese Aufteilung verschiebt für die Feedbackbedingungen logischerweise das Bild auf der deskriptiven Ebene. Die Feedbackgruppen haben im Vergleich zur Kontrollgruppe im Mittel eher weniger Zeit auf die Beantwortung der Erstantworten verwendet. Die Zeit, die sie insgesamt für das Experiment mehr benötigen als die Kontrollgruppe, stammt deshalb im Wesentlichen aus der Bearbeitung der Zweitantworten.

Die varianzanalytische Auswertung der Werte ergibt, dass sich die experimentellen Bedingungen weder hinsichtlich der Bearbeitungszeiten der Erstantworten (F (4, 485) = 1.93; $p > .05$; $\eta^2 = .02$) noch der Zweitantworten (F (3, 392) = 0.65; $p > .05$; $\eta^2 = .01$) statistisch bedeutsam voneinander unterscheiden.

Tabelle 26 Deskriptive Statistik der Bearbeitungszeiten in der Treatmentphase (N = 490)

		Anzahl aufgewendeter Minuten im Experiment						
		Gesamt			Erstantworten (34 Items)		Zweitantworten	
	N	*M*	*SD*	*Med*	*M*	*SD*	*M*	*SD*
Kein Feedback	94	34.37	10.54	34.15	34.37	10.54	–	–
Knowledge of Result	95	37.44	11.95	36.05	32.70	10.17	4.74	3.46
Metakognitiver Prompt	109	35.92	9.92	33.57	31.61	8.59	4.30	2.39
Fehlererklärung	94	36.40	8.47	36.18	31.71	7.40	4.68	2.19
Inferenzprompt	98	35.79	11.09	35.11	31.00	9.40	4.79	3.13

In einem weiteren Schritt wurde für die Feedbackbedingungen die Gesamtbearbeitungszeit der Zweitantworten pro Person an der Anzahl der benötigten zweiten Versuche relativiert. Daraus resultiert die Zeit (Sekunden), die pro Person im Durchschnitt für eine Zweitantwort aufgewendet wurde. Anhand dieses Wertes wurde wiederum die mittlere Gruppenleistung berechnet und verglichen.

Äquivalent dazu bietet sich für Vergleiche zwischen Zweit- und Erstantworten an, auch die Bearbeitungszeiten für Erstantworten auf der Itemebene zu betrachten. Dabei ist allerdings zu berücksichtigen, dass mit der ersten Präsentation eines neuen Textes gleichzeitig auch das erste Item präsentiert wurde. Da bei der Bearbeitung der Units eher erwartet werden kann, dass zunächst der Text vollständig gelesen wurde und dann die Items beantwortet wurden, fällt die Bearbeitungszeit für das erste Item sehr wahrscheinlich mit der Textrezeption zusammen, wodurch für das erste Item einer jeden Unit verhältnismäßig mehr Zeit aufgewendet worden sein sollte als für die ersten Versuche der restlichen Items. Demzufolge sind für die Analyse der Zeiten für eine durchschnittliche Erstantwort die jeweils ersten Items einer Unit getrennt ausgewertet.

Die deskriptiven Kennwerte für die Erstantworten und die Zweitantworten sind in Tabelle 27 dargestellt.

Die RM-ANOVA für die Bearbeitungszeiten für eine Erstantwort, und zwar bezogen auf erste Items der Units einerseits und alle weiteren Items andererseits, bringt einen hoch signifikanten Effekt des Innersubjektfaktors hervor (F (1, 485) = 1825.12; $p < .001$; partielles $\eta^2 = .79$). Statistisch bedeutsame Unterschiede zwischen den Versuchsgruppen (F (4, 485) = 0.58; $p > .05$; partielles $\eta^2 = .01$) oder Wechselwirkungen zwischen Versuchsgruppen und der aufgewendeten Zeit in den beiden verschiedenen Maßen für Erstantworten (F (4, 485) = 0.13; $p > .05$; partielles $\eta^2 = .00$) sind nicht nachweisbar. Das Ergebnis spricht also dafür, dass beim ersten Präsentieren eines Textes (zusammen mit dem ersten Item) dieser eher erst vollständig gelesen wurde.

Tabelle 27 enthält neben den Bearbeitungszeiten für eine durchschnittliche Erstantwort auch die entsprechenden Zeiten für eine durchschnittliche Zweitantwort. Im Vergleich der deskriptiven Kennwerte wird dabei ersichtlich, dass in den Feedbackbedingungen für einen Zweitversuch weniger Zeit aufgewendet wird als für einen Erstversuch. Die RM-ANOVA belegt dies durch einen hoch signifikanten Effekt für den Innersubjektfaktor, das bedeutet den Vergleich zwischen Zeit für Erstantwort versus Zweitantwort (F (1, 392) = 1750.09; $p < .001$; partielles $\eta^2 = .82$). Der Zwischensubjektfaktor ist dagegen nicht

signifikant (F (3, 392) = .54; p > .05; partielles η^2 = .00), ebenso nicht der Interaktionseffekt (F (3, 392) = 1.09; p > .05; partielles η^2 = .01).

Tabelle 27 Bearbeitungszeiten auf Itemebene

		Zeit für eine durchschnittliche Antwort (in Sekunden)					
		Erstantwort				Zweitantwort	
		1. Items der Units (eher Textrezeption)		Alle Items ohne die 1. Items pro Unit			
	N	*M*	*SD*	*M*	*SD*	*M*	*SD*
Kein Feedback	94	167.39	71.07	42.76	16.79	–	–
Knowledge of Result	95	164.60	74.30	39.20	13.21	14.37	10.06
Metakognitiver Prompt	109	164.14	58.91	37.38	12.35	13.51	7.49
Fehlererklärung	94	159.19	51.96	38.37	11.70	14.83	7.56
Inferenzprompt	98	159.83	71.05	36.76	11.81	14.79	7.29

Anmerkungen. – = nicht verfügbar (Kontrollgruppe ohne Zweitversuche).

Die Korrelation der durchschnittlichen Bearbeitungszeit einer Aufgabe im ersten Versuch und im zweiten Versuch ist mit r = .46 (p < .001) statistisch bedeutsam. Zusammenhänge zwischen der Gesamtbearbeitungsdauer für Erst- bzw. Zweitantworten und den entsprechenden Leistungen sind nicht nachweisbar. Die Korrelation der Bearbeitungszeit für Erstantworten und die Leistung in den Erstantworten ergibt r = - .07 (p > .05). Ebenso ist zwischen der Bearbeitungszeit für die Zweitantworten und die relative Lösungshäufigkeit in den Zweitversuchen mit r = - .09 (p > .05) kein Zusammenhang nachweisbar.

Posttest

Hinsichtlich der Zeit, die für den Posttest aufgebracht wurde, unterscheiden sich die Versuchsgruppen nicht in statistisch bedeutsamer Weise voneinander (F (4, 357) = 1.67; p > .05; η^2 = .02). Die deskriptiven Kennwerte sind in Tabelle 28 zusammengefasst.

In Abbildung 11 sind die durchschnittlichen Bearbeitungszeiten für die Erst- und die Zweitantworten der Treatmentphase sowie die Bearbeitungszeit für den Posttest zur Illustrierung der Daten gemeinsam dargestellt.

Tabelle 28 Deskriptive Statistik der Bearbeitungsdauer des Posttests (N = 362)

	Anzahl aufgewendeter Minuten im Posttest (13 Items)			
	N	*M*	*SD*	*Med*
Kein Feedback	75	9.10	3.34	8.92
Knowledge of Result	71	8.40	3.64	7.96
Metakognitiver Prompt	80	8.20	3.25	8.62
Fehlererklärung	60	8.24	3.18	8.51
Inferenzprompt	76	7.70	3.43	7.54

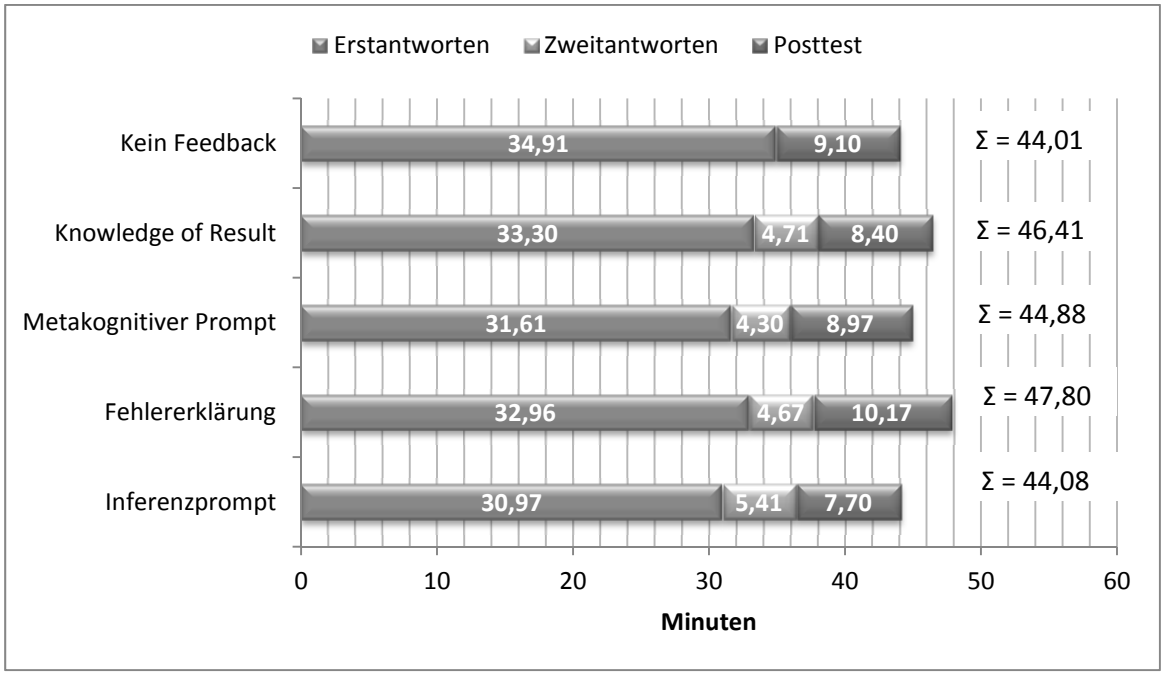

Abbildung 11 Darstellung der mittleren Bearbeitungszeiten von Experiment und Posttest.

Die Analysen der Bearbeitungszeiten im Experiment und Posttest zusammengenommen ist festzuhalten, dass die Feedbackinterventionen nicht zu einer statistisch bedeutsamen längeren Beschäftigung bzw. Auseinandersetzung mit den Aufgaben geführt hat. Die Zeit, die die Feedbackbedingungen insgesamt länger für die Experimentalphase aufgebracht haben, resultiert aus der Zeit, die durch die Beantwortung der Zweitversuche nötig wurde. Hierbei unterscheiden sich die (einfache und die elaborierten) Feedbackbedingungen aber ebenfalls nicht statistisch bedeutsam voneinander. Die Auswertungen der durchschnittlichen Bearbeitungszeit für eine Erstantwort im Vergleich zu einer Zweitantwort ergänzen das Befundlage insofern, dass hier gezeigt werden

konnte, dass die Zeit, die in den Feedbackgruppen für eine Zweitantwort aufgewendet wurde, signifikant kürzer ausfällt als für eine Erstantwort.

6.5 Haupteffekt von Feedback auf Testangst

Mögliche Einflüsse der Feedbackinterventionen auf die Testangst der Probanden sind für die zwei Subskalen Besorgtheit und Aufgeregtheit mittels RM-ANOVAs mit den Versuchsgruppen als Zwischensubjektfaktor und dem Erhebungszeitpunkt der Testangstskala (vor vs. nach dem Experiment) als Innersubjektfaktor überprüft. Die deskriptiven Kennwerte sind Tabelle 29 zu entnehmen. Abbildung 12 illustriert die durchschnittlichen Ausprägungen der Besorgtheit und Aufgeregtheit vor und nach dem Experiment.

Tabelle 29 Deskriptive Statistik der Subskalen zur Testangst (N = 495)

		Testangst							
		Subskala Besorgtheit				Subskala Aufgeregtheit			
		Prä		Post		Prä		Post	
	N	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Kein Feedback	95	2.29	0.67	2.28	0.75	1.61	0.62	1.71	0.74
Knowledge of Result	96	2.30	0.67	2.33	0.77	1.68	0.71	1.84	0.82
Metakognitiver Prompt	109	2.32	0.70	2.34	0.79	1.65	0.56	1.59	0.65
Fehlererklärung	96	2.34	0.70	2.33	0.74	1.69	0.68	1.70	0.72
Inferenzprompt	99	2.34	0.68	2.33	0.86	1.62	0.56	1.65	0.73

Bezüglich der RM-ANOVA der Komponente Besorgtheit ist kein signifikanter Haupteffekt auf den Innersubjektfaktor (F (1, 490) = .03; $p > .05$; partielles $\eta^2 = .00$) oder den Zwischensubjektfaktor (F (4, 490) = .10; $p > .05$; partielles $\eta^2 = .00$) sowie kein signifikanter Interaktionseffekt beider (F (4, 490) = .13; $p > .05$; partielles $\eta^2 = .00$) nachweisbar. Ein analoges Ergebnis bringt die RM-ANOVA für die Komponente Aufgeregtheit hervor: für den Innersubjektfaktor ist ebenso kein signifikanter Haupteffekt nachweisbar (F (1, 490) = 3.30; $p > .05$; partielles $\eta^2 = .01$) wie für den Zwischensubjektfaktor (F (4, 490) = .85; $p > .05$; partielles $\eta^2 = .01$) und auch der Interaktionseffekt zwischen Messzeitpunkt und Versuchsgruppen ist nicht signifikant (F (4, 490) = 2.13; $p > .05$; partielles $\eta^2 = .02$).

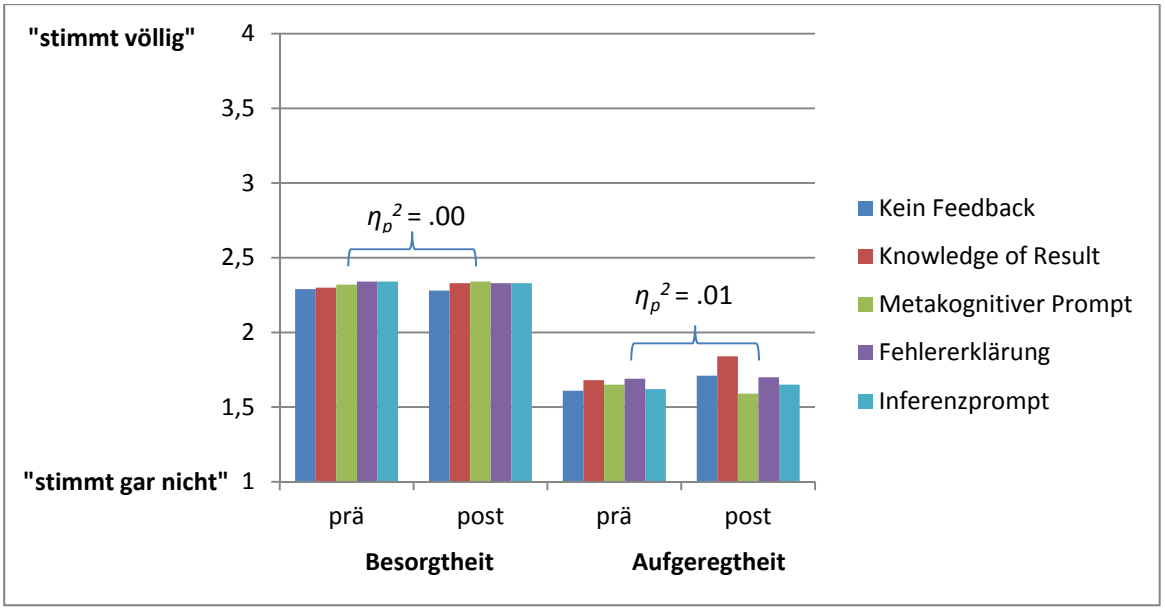

Abbildung 12 Ausprägungen in den Subskalen zur Testangst.

Anmerkungen. Skalierung von 1 für „stimmt gar nicht“ bis 4 für „stimmt völlig“.

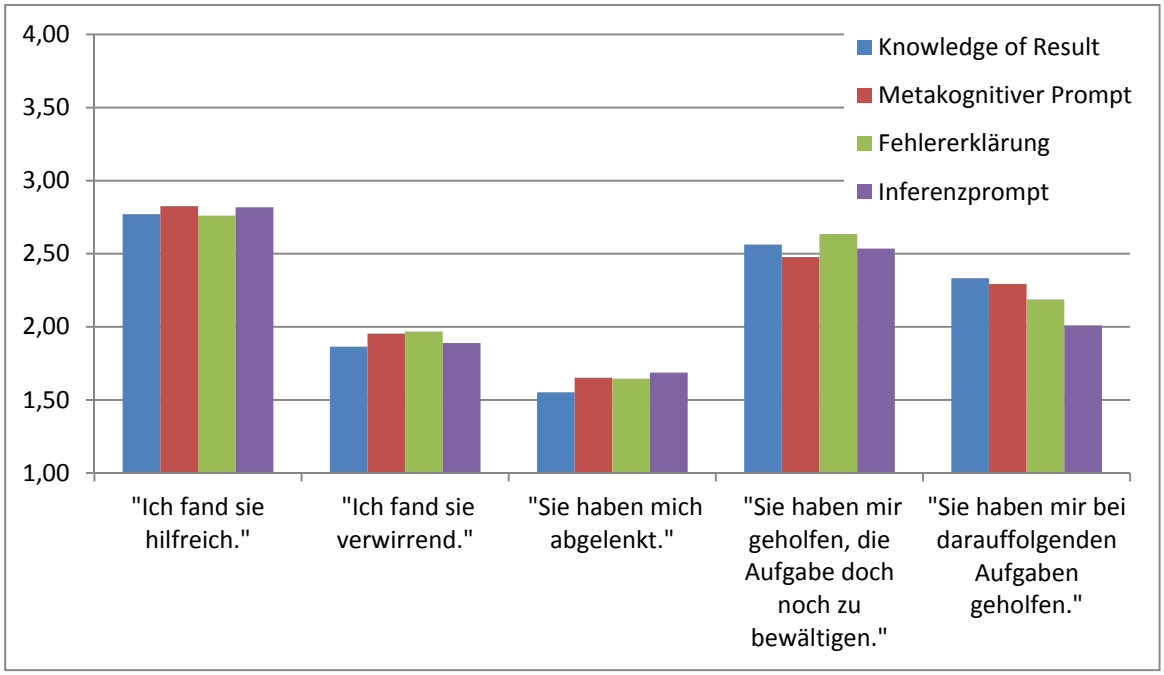

Abbildung 13 Gruppenmittelwerte (Einzelitems) zur Einschätzung der Feedbacks.

Anmerkungen. Skalierung von 1 für „stimmt gar nicht“ bis 4 für „stimmt völlig“, die abgebildeten Mittelwerte von Item 2 und Item 3 (von links) basieren auf den nicht rekodierten Rohwerten.

6.6 Analyse der wahrgenommenen Nützlichkeit der Feedbacks

Die Einschätzung der Feedbackgruppen bezüglich der Nützlichkeit der erhaltenen Rückmeldungen fällt, wie die deskriptiven Kennwerte in Tabelle 30 zeigen, sehr ähnlich aus. Die ANOVA ergibt dementsprechend keinen signifikanten Haupteffekt des Feedbacks auf die Einschätzung der Nützlichkeit der Rückmeldungen durch die Feedbackempfänger (F (3, 396) = 0.41; $p > .05$; $\eta^2 = .00$). Bei der Skalierung der Items von 1 (keine Zustimmung) bis 4 (uneingeschränkte Zustimmung) spricht der Gesamtgruppenmittelwert von ungefähr M = 2.80 insgesamt für eine eher positive Einschätzung der Nützlichkeit der Rückmeldungen.

In Abbildung 13 sind die Gruppenmittelwerte für die 5 Items des Fragebogens zur Nützlichkeit separat aufgeführt. Dabei ist zu beachten, dass das zweite und dritte Item, die für die Skalenbildung rekodiert wurden, in der Abbildung die nicht rekodierten Werte widerspiegeln.

Tabelle 30 Deskriptive Statistik zu Einschätzungen der Feedbacks (N = 400)

	Wahrgenommene Nützlichkeit der Rückmeldungen [a]			
	N	*M*	*SD*	*Med*
Knowledge of Result	96	2.85	0.64	3.00
Metakognitiver Prompt	109	2.80	0.58	2.80
Fehlererklärung	96	2.79	0.64	2.80
Inferenzprompt	99	2.76	0.50	2.80

Anmerkungen. [a] Skalierung von 1 für „stimmt gar nicht“ bis 4 für „stimmt völlig“.

Zusammenhänge zwischen der Einschätzung der Nützlichkeit der Rückmeldungen und der Leistung bestehen nicht, weder bezüglich der Erstantworten (r = - .04; $p > .05$), der Zweitantworten (r = .08; $p > .05$) oder der Leistung im Posttest (r = .06; $p > .05$).

6.7 Zusammenfassung der zentralen Ergebnisse

Die für die Beantwortung der Fragestellungen zentralen Ergebnisse können wie folgt zusammengefasst werden. Bezüglich der Auswirkungen auf *das Textverständnis bzw. die Lesekompetenz* zeigte sich, dass:

- keine der elaborierten Feedbackinterventionen entgegen den a-priori formulierten Annahmen einen statistisch signifikanten Effekt auf die Leistung hervorbrachte, weder in der Treatmentphase (Erst- bzw. Zweitantworten) noch im anschließenden Posttest oder im Follow-up,
- die einfache Feedbackart Knowledge of Result erwartungskonform keinen Effekt auf die Leistung hatte.

Außerdem ist aus den Leistungsanalysen hervorzuheben, dass die Anzahl der „Treatmentmomente", also die Häufigkeit der Feedbackgaben, in allen Feedbackgruppen durchschnittlich etwa 20 Rückmeldungen (bei 34 Items) betrug.

Aus den Analysen der *Bearbeitungszeiten* ist zusammenfassend festzuhalten, dass

- die elaborierten Feedbackinterventionen nicht zu höheren Bearbeitungszeiten im Treatment geführt haben,
- die Feedbackgruppen signifikant sehr viel weniger Zeit für eine Zweitantwort aufwendeten als für eine Erstantwort.

Ferner belegen die Analysen, dass

- die Feedbackinterventionen im Allgemeinen keine nachteiligen Auswirkungen auf die *Aufgeregtheit oder Besorgtheit* (Testangst) der Probanden bezüglich ihrer Leistungen im Experiment hatten,
- die subjektiven Einschätzungen der *Nützlichkeit der Rückmeldungen* im Durchschnitt eher positiv ausfielen,
- sich die Feedbackbedingungen in dem Ausmaß der wahrgenommenen Nützlichkeit der Hilfen nicht unterschieden.

7 Diskussion

Als Rahmen der Interpretation und Diskussion der zuvor berichteten Untersuchungsergebnisse werden zunächst die wichtigsten Eckpunkte des theoretischen und methodischen Konzepts des Experiments wiederholt: Das Anliegen des vorliegenden Experiments bestand darin, die Wirksamkeit verschiedener elaborierter Feedbackinterventionen auf das Textverständnis bzw. die Lesekompetenz zu untersuchen. Umgesetzt wurde das Treatment in einem Setting, welches Ausdruck der notwendigen Kompatibilität des Experiments mit dem dynamischen Testgedanken ist (vgl. Abschnitt 3.5.3). Der Untersuchungskontext zeichnet sich im Wesentlichen durch das in einer einmaligen Sitzung durchgeführte, computerbasierte „Test-Lern-Setting“ aus, in dem Feedback unmittelbar auf Antworten im Test gegeben wurde und ansonsten keine weiteren Unterstützungsmaßnahmen implementiert waren.

Als Treatmentbedingungen wurden drei elaborierte Feedbackarten umgesetzt, die vor dem Hintergrund der Rahmenbedingungen des Untersuchungskontextes und unter Berücksichtigung empirischer Befunde zu Feedback und zum Textverstehen abgeleitet wurden. Diese Feedbackarten sind a) Fehlererklärung, b) ein metakognitiver Prompt zur Anregung der Verstehensüberwachung sowie c) Inferenzprompts, die kognitive Hinweise zum Herstellen der durch eine Aufgabenstellung geforderten Inferenz darstellen. Daneben wurden als Kontrollbedingungen die einfachste Feedbackart (Knowledge of Result ohne weitere Hinweise) sowie eine feedbackfreie Kontrollbedingung implementiert.

Die feedbackfreie Kontrollbedingung hat nicht nur kein Feedback erhalten, sondern auch alle Aufgaben nur einmal beantwortet. Dadurch entspricht diese Bedingung einer konventionellen Testbearbeitung. Die Kontrollbedingung Knowledge of Result wurde umgesetzt, weil über sie eine angemessenere Bewertung der potentiellen Effekte der elaborierten Feedbackarten möglich ist. Vom Feedback Knowledge of Result selbst sind keine Effekte auf das Textverstehen/die Lesekompetenz zu erwarten, aber die Probanden dieser Bedingung arbeiten aufgrund der identischen Feedbackprozedur unter denselben Bedingungen wie die Probanden der elaborierten Feedbackinterventionen. Dieser Aspekt wird durch eine feedbackfreie Kontrollgruppe nicht abgedeckt.

Für die Überprüfung der Feedbackeffekte wurden nicht nur, wie üblich, Transfertests (Posttest und Follow-up) genutzt, sondern auch die Leistung in der Treatmentphase. Diese ist in Erstantworten und Zweitantworten unterteilt. Die Zweitantworten reflektieren die Korrekturleistung nach der Feedbackgabe und die Erstantworten sind auch als Transferleistung interpretierbar (vgl. Abschnitt 3.4). Durch die zusätzliche Berücksichtigung dieser beiden Maße ist die Wirkung der Feedbacks auf die Leistung facettenreicher zu beurteilen.
Neben dem Fokus der Interventionseffekte auf das Textverständnis/die Lesekompetenz sind auch die Effekte der Feedbackinterventionen auf die Bearbeitungs- bzw. Verweilzeiten für die Testaufgaben (nach Feedbackgabe) untersucht worden. Eine untergeordnete Rolle spielen die Analysen möglicher Feedbackeffekte auf die Testangst und die Wahrnehmung der Nützlichkeit der Hilfen.

Die nachfolgende Diskussion der Ergebnisse fokussiert zunächst die Auswirkungen der Feedbackinterventionen auf das Textverständnis/die Lesekompetenz und integriert dabei die Erkenntnisse bezüglich der Bearbeitungszeiten sowie der Auswirkungen hinsichtlich der Testangst und der wahrgenommenen Nützlichkeit der Rückmeldungen. Im Anschluss daran werden darauf abhebend übergreifende Aspekte der Wirksamkeit bzw. Wirklosigkeit der Feedbackinterventionen erörtert. Danach folgt die kritische Auseinandersetzung mit der Methodik der Untersuchung. Die Diskussion schließt mit einem Gesamtfazit und der Überleitung zum zweiten Experiment.

7.1 Über die Auswirkungen der Feedbackinterventionen auf das Textverständnis/die Lesekompetenz

Das zentrale Ergebnis des Experiments ist, dass die elaborierten Feedbackinterventionen entgegen der a-priori Hypothesen keine statistisch bedeutsamen Leistungssteigerungen nach sich ziehen konnten. Keine der Feedbackarten – weder die Fehlererklärung noch der metakognitive Prompt oder Inferenzprompt – hat sich für die Leser im Allgemeinen als nützlich bzw. nutzbar erwiesen.
Wie im theoretischen Teil der Arbeit bezüglich Verständnisschwierigkeiten dargelegt wurde, sind die Möglichkeiten und Gründe für misslingendes Verstehen mindestens so vielfältig und komplex wie die Anforderungen des erfolgreichen Textverstehens. Für die

Wirkung oder eben die ausbleibende Wirkung von Feedbackinterventionen kann wohl dasselbe gelten, insbesondere wenn es für das Textverstehen gegeben wird. Die Feedbackliteratur bietet allerdings keine Modelle zur Erklärung der Wirkungslosigkeit von (elaboriertem) Feedback (Kluger & DeNisi, 1996). In der Regel werden die Effekte von Feedbackarten primär auch auf deren Inhalt zurückgeführt.

Die Diskussion der Ergebnisse der vorliegenden Arbeit erschöpft sich nicht darin, zu schlussfolgern, dass die gewählten Feedbackinhalte ungeeignet für das Textverstehen sind. Stattdessen werden unter Einbezug des Untersuchungskontexts und der kognitiven und motivationalen Anforderungen an den Lerner mehrere mögliche Ursachen des Ausbleibens der Effekte in den elaborierten Feedbackinterventionen erörtert.

Erstantworten in der Treatmentphase

Der Verwendung der Erstantworten der Treatmentphase zur Beurteilung der Feedbackwirksamkeit liegt der folgende Gedanke zugrunde: wenn eine Feedbackart geeignet ist, um aus ihr neue Erkenntnisse zu gewinnen und diese auf neue Aufgabenstellungen transferieren zu können, sollte sich ein Effekt nicht erst in einem anschließenden, konventionellen Transfertest (Posttest ohne Feedback) zeigen, sondern – in dem Test-Lern-Setting dieser Untersuchung – auch schon für ähnliche Aufgaben im Verlauf der Test- bzw. Treatmentphase.

Die Analysen ergaben, dass sich die Versuchsgruppen in den Erstantworten der Treatmentphase nicht signifikant voneinander unterschieden. Damit wurde die a-priori formulierte Erwartung, dass Interventionen mit den drei umgesetzten elaborierten Feedbackarten – Fehlererklärung, metakognitiver Prompt und Inferenzprompt – zu einer Verbesserung der Leistung in den ersten Antwortversuchen der Treatmentphase führen, nicht erfüllt. Das Ergebnis bestätigt dagegen die Hypothese bezüglich des einfachen Feedbacks Knowledge of Result, das sich als nicht geeignet erwies, um die Leistung in den nachfolgender Aufgaben zu verbessern.

In den Überlegungen zu möglichen Gründen für das Ergebnis der elaborierten Feedbackinterventionen ist zunächst in Betracht zu ziehen, dass das Ausbleiben der Effekte nicht zwangsläufig auf ausgebliebene Lerneffekte zurückzuführen ist. Es ist ebenso möglich, dass die Probanden im Allgemeinen aus den Rückmeldungen lernten, dieses Wissen in den Erstantworten der Treatmentphase aber (noch) nicht realisieren

konnten. Dieser mögliche „verdeckte“ Lerneffekt erscheint insbesondere vor dem Hintergrund des Charakters der Intervention einer Erörterung wert. Dabei werden vor allem zwei Aspekten des Untersuchungskontexts als relevant erachtet.

Ein erster möglicher Einfluss wird darin gesehen, dass die Testsituation in der Experimentalphase für die Probanden der Feedbackbedingungen durch Unterbrechungen und wiederholtes Bearbeiten von Aufgaben gekennzeichnet ist. Das impliziert nicht nur einen höheren Bearbeitungsaufwand, sondern dem Probanden werden wiederholt die Unzulänglichkeiten seiner Leistung und möglicherweise bestehende Fehleinschätzungen der eigenen Leistungsfähigkeit vor Augen geführt. Die Experimentalsituation dürfte daher von den meisten Probanden der Feedbackgruppen mindestens als ungewohnt, möglicherweise auch als verunsichernd oder störend empfunden worden sein. Dabei ist auch zu berücksichtigen, dass in den ersten Versuchen mit durchschnittlich etwa 20 falschen Antworten (von 34 Items) im Allgemeinen relativ viele Fehler gemacht wurden und entsprechend häufig Feedback gegeben wurde.
Die Situation, unter der die Bearbeitung der Erstantworten in diesem Experiment erfolgte, ist also per se eine andere als in einem interventionsfreien Posttest, der normalerweise herangezogen wird, um die Wirksamkeit einer Feedbackintervention zu beurteilen. Auch die a-priori formulierte Annahme eines positiven Effekts der elaborierten Feedbacks in den Erstantworten wurde in erster Linie aus Feedbackstudien abgeleitet, deren Befunde in der Regel auf Transferleistungen in einem Posttest beruhen (vgl. Abschnitt 3.5.1.1 und Tabelle 4). Insofern ließe sich die Diskrepanz zwischen der Hypothese und den ermittelten Ergebnissen nachvollziehen.

Als ein weiterer Grund der Ineffektivität der elaborierten Feedbackinterventionen in den Erstantworten wird auch in Betracht gezogen, dass sich die Probanden auf die Feedbackgabe verlassen haben und nicht die Notwendigkeit sahen, die Informationen der Rückmeldungen auch auf nachfolgende Aufgaben zu übertragen. Eine ähnliche Vermutung wurde auch schon von Shute (2008) im Zusammenhang mit dem Geben unmittelbarer Rückmeldungen formuliert (vgl. Abschnitt 3.5.1.2). Sie weist darauf hin, dass die unmittelbare und damit zuverlässige Feedbackgabe dazu führen kann, dass sich der Lerner auf die Hilfestellungen verlässt, deshalb weniger sorgfältig bzw. selbstständig arbeitet und somit keine Verbesserung seiner Leistung erfahren kann.

Die Schlussfolgerung von Shute, dass das sich Verlassen auf die Rückmeldungen das Lernen unterbindet, ist für dieses Experiment nicht zwangsläufig zu ziehen. Es ist durchaus denkbar, dass sich die Schüler aufgrund der besonderen Bearbeitungssituation mit Feedback schlicht nicht veranlasst sahen, darüber nachzudenken, ob die Rückmeldungen auch auf nachfolgende Aufgaben übertragbar sind, und/oder entsprechende Anstrengungen zu unternehmen. Dieser Eindruck mag durch die zweiten Antwortversuche, die als eine Art Rückversicherung oder zweite Chance gesehen worden sein könnten, noch verstärkt worden sein. Nichtsdestotrotz könnten die Informationen und Hinweise, die in den Rückmeldungen vermittelt wurden, gespeichert und später im Posttest, wenn das „Stützgerüst" der Feedbackintervention wegfällt, dennoch angewendet werden.

Beide angesprochenen Merkmale der Interventionssituation könnten, einzeln oder gemeinsam, bewirkt haben, dass erworbenes Wissen aus den Rückmeldungen (noch) nicht im (Antwort-)Verhalten realisiert wurde. Aber diese Umgebungsbedingungen könnten ebenso dazu geführt haben, dass kein Lerneffekt eintrat, vor allem dann, wenn sie die Motivation bzw. die Bemühungen der Probanden für eine Umsetzung der Rückmeldungen untergruben oder abbrachen.
Weiterhin besteht die Möglichkeit, dass der Lerneffekt in den elaborierten Interventionen ausblieb, weil die Feedbackinhalte für das Textverstehen unbrauchbar waren. Das würde bedeuten, dass weder die Fehlererklärung noch der metakognitiver Prompt zur Anregung der Verstehensüberwachung oder die Prompts zum Herstellen geforderter Inferenzen geeignet waren, um von den Probanden noch in der Treatmentsituation, also sozusagen unmittelbar, auf nachfolgende Aufgaben übertragen werden zu können.

Welche der aufgeführten möglichen Erklärungen für die ausgebliebenen Effekten der elaborierten Feedbackinterventionen in den Erstantworten verantwortlich ist, kann nicht mit Sicherheit bestimmt werden. Hinsichtlich der Erörterungen, ob es sich dabei um einen tatsächlich ausgebliebenen oder um einen verdeckten Lerneffekt handelt, bekräftigen die Ergebnisse der Leistungen in den Zweitantworten jedoch die These, dass durch die elaborierten Feedbacks kein Lerneffekt eingetreten ist.

Zweitantworten in der Treatmentphase

Die Leistung in den zweiten Antwortversuchen spiegelt die Wirksamkeit einer Feedbackart bezüglich der Korrektur einer falschen Erstantwort wider. Das Interessante an den Zweitantworten ist, dass sie die direkte Wirkung einer Rückmeldung abbilden. Die anderen Maße, die Erstversuche und die Posttests, setzen immer einen Transfer voraus, der insbesondere bei kognitiven anspruchsvollen Fähigkeitsbereichen wie dem Textverstehen zusätzliche Anforderungen an den Lerner stellt. Aber die Bearbeitung einer Aufgabe, nachdem für diese eine Hilfestellung gegeben wurde, ist die unmittelbare Nützlichkeit eines Feedbacks. Wenn ein Feedback nicht für die Korrektur der Aufgaben genutzt werden kann, für die es gegeben wurde, verspricht es eigentlich auch kein Potential für Transferleistungen zu haben.

Die Analysen der Zweitantworten zeigten, dass sich die drei Bedingungen elaborierten Feedbacks in ihren Korrekturleistungen falscher Antworten nicht bedeutsam von der Bedingung Knowledge of Result abhoben. Das bedeutet, dass die elaborierten Informationen gegenüber dem Knowledge of Result keinen Mehrwert für die Korrektur falscher Antworten aufwiesen. Dieses Ergebnis widerspricht den Erwartungen und bedeutet, dass die Inhalte der elaborierten Feedbacks nicht nützlich oder nutzbar waren. Damit sind bewusst zwei verschiedene Aspekte impliziert: die Feedbackinhalte könnten zum einen tatsächlich ungeeignet für die Korrektur des Textverständnisses gewesen sein und zum anderen könnten sie (z.B. durch die Anforderungen, die sie dennoch an den Leser stellen) gemieden worden sein. Diese beiden Aspekte werden an späterer Stelle nach der Diskussion der weiteren Ergebnisse erläutert, da die Erklärungen dann nach der Betrachtung aller Resultate gesammelt dargestellt werden können. Alternative Erklärungen für die ausgebliebenen Effekte in den Zweitantworten sind nicht plausibel.
Auch die Ergebnisse zur Testangst geben keinen Anlass: zum einen waren die Probanden der Feedbackbedingungen vor Beginn des Experiments nicht aufgeregter oder besorgter bezüglich der anstehenden Testsituation mit Feedbackgabe als die Kontrollgruppe, die auf einen „normalen“ Test ohne Feedbackgabe vorbereitet war. Zum anderen wirkten sich die Feedbackinterventionen auch nicht nachteilig auf das Befinden nach Abschluss des Experiments aus.
Auch die Analyse der Bearbeitungszeiten für die Zweitantworten geben keine weiteren Anhaltspunkte. Die vier Feedbackgruppen sind sich hinsichtlich der aufgewendete Zeit hoch ähnlich, ist gibt keine bedeutsamen Abweichungen zwischen den Gruppen.

Inwiefern die Verweilzeiten in den zweiten Antworten für die Art oder die Extensität der Auseinandersetzung mit den Rückmeldungen und den wiederholten Antwortversuchen sprechen, kann hinterfragt werden. Im Durchschnitt wurden für eine Zweitantwort ungefähr 14 Sekunden aufgewendet, diese Zeit schließt das Lesen der Rückmeldung und die Aktivitäten bis zur Abgabe der erneuten Antwort ein. Im Vergleich dazu wurde für eine Erstantwort im Durchschnitt das Dreifache der Zeit aufgewendet. Ob der Durchschnittswert von etwa 14 Sekunden für eine Zweitantwort nun aber für oder gegen die Nutzung der Rückmeldungen spricht, ist kaum auszumachen. Zumindest stellt sich die Situation so dar, dass die zweiten Versuche wohl nicht ignoriert und sozusagen „weggeklickt“ wurden.

Unabhängig von den nicht gefundenen Gruppenunterschieden erscheint es bemerkenswert, dass die durchschnittliche Korrekturleistung aller vier Feedbackgruppen bei ungefähr 44 % lag. Praktisch bedeutet das, dass annähernd jeder zweite Fehler eines Erstversuchs im zweiten Versuch richtig gelöst werden konnte. Dieses Ergebnis wird dadurch unterstrichen, dass die Fehlerquote in den Erstantworten bei allen Gruppen (also die feedbackfreie Kontrollgruppe eingeschlossen) mit durchschnittlich 20 Fehlern bei 34 Aufgaben (58.82 %) verhältnismäßig hoch war und damit eher viele Zweitversuche durchgeführt wurden.

Es stellt sich die Frage, inwiefern allein das Rückmelden von Fehlern bei gleichzeitiger Einräumung einer zweiten Antwortgelegenheit schon einen Nutzen darstellt. Ein Effekt nur durch Knowledge of Result wäre dadurch gegeben, dass allein die Botschaft, dass eine Antwort falsch ist, zu (meta-)kognitiven Prozessen führt, dem Einsatz von Lesestrategien etwa, die in einem verbessertem Textverständnis und ergo der erfolgreichen Zweitantwort resultieren. Allerdings ist auf der Grundlage der Feedbackliteratur davon auszugehen, dass genau dieser Effekt höchst unwahrscheinlich für hierarchiehöhere Prozesse der Lesekompetenz ist, und dementsprechend fiel die Hypothese für diese Arbeit aus.

Immerhin ist auch zu bedenken, dass das wiederholte Beantworten ein und derselben Aufgabe zwangsläufig die Antwortbedingungen verändert. Die Ratewahrscheinlichkeit bei vier Antwortalternativen, die im zweiten Versuch noch zur Auswahl standen, läge bei einer herkömmlichen Testbearbeitung und annähernd gleich plausiblen Antwortalternativen bei 25 %. Demgegenüber erscheint die tatsächliche Korrekturleistung von durchschnittlich etwa 44 % zunächst deutlich höher. Aber der

Vergleich ist so wohl nur bedingt zulässig und verleitet möglicherweise eher zu einer Überinterpretation. Denn das wiederholte Beantworten ein und derselben Aufgabenstellung erfolgt unter veränderten (besseren) Bedingungen als eine Erstantwort.
Eine Schlussfolgerung für die Bedeutung der Ergebnisse dieses Experiments kann nicht gezogen werden. Der Aspekt an sich wird in der Gesamtdiskussion aber nochmals aufgegriffen und in Hinblick auf die Bedeutung für Feedback und/oder eines dynamischen Kurzzeitlerntests diskutiert.

Zwischenfazit: Das Fazit aus den Ergebnissen der Zweitantworten ist, dass die elaborierten Feedbackinformationen keinen Mehrwert für die Korrekturleistung nach sich gezogen haben. Im Zusammenspiel mit den Ergebnissen der Erstantworten legt das, trotz der für die Erstantworten eingeräumten Möglichkeit eines verdeckten Lerneffekts, die Schlussfolgerung nah, dass die elaborierten Feedbackarten, entgegen der Annahmen, im Allgemeinen keine Lerneffekte bewirkt hatten.

Leistung im Posttest

Der Posttest, der in dieser Untersuchung mit kurzer Unterbrechung unmittelbar nach der Treatmentphase administriert wurde, bildet die Transferleistung der Feedbackgruppen nach der Intervention ab. Die Resultate des Posttest zeigen, dass es keine statistisch bedeutsamen Unterschiede zwischen den Versuchsgruppen gab. Keine der elaborierten Feedbackinterventionen konnte hier gegenüber den Kontrollbedingungen eine Leistungssteigerung erzielen. Unter Berücksichtigung der Ergebnisse der Treatmentphase ist davon auszugehen, dass im Posttest keine Leistungssteigerungen möglich waren, weil die Treatments mit elaborierten Feedbacks keinen (ausreichenden) Lerneffekt bewirkten.

Darüber hinaus deutet der Vergleich der Leistungen in der Treatmentphase und dem Posttest für die Bedingung Inferenzprompt einen Leistungsabfall zum Posttest hin an. Dabei fällt der Unterschied berechnet anhand des Differenzwertes zwischen beiden Testteilen signifikant aus. Die Berechnungen anhand der ANOVA mit Messwiederholungen ergaben dagegen keine signifikante Verschlechterung in der Gruppe Inferenzprompt. Natürlich ist zu berücksichtigen, dass die Feedbackbedingungen durch die Feedbackgaben und die wiederholten Antworten im Allgemeinen eine nicht

unbeachtliche Mehrarbeit im Vergleich zur feedbackfreien Kontrollgruppe geleistet hatten. In den Bedingungen mit elaborierten Rückmeldungen kommt im Vergleich zu Knowledge of Result zumindest theoretisch noch die Verarbeitung längerer Mitteilungen hinzu. Das könnte durchaus dazu führen, dass Ermüdungseffekte stärker/früher auftreten oder dass die Testmotivation eher abnimmt. Aber es ist nicht auszumachen, warum diese möglichen Einflüsse auf die Gruppe mit den Inferenzprompts zutreffen sollte, dagegen aber nicht auf die Gruppen mit den anderen elaborierten Feedbacks. Auch die Analyse der Bearbeitungszeiten des Posttests gibt keinen entsprechenden Anhaltspunkt. Die Versuchsgruppen weichen diesbezüglich nicht statistisch bedeutsam voneinander ab.

Allen Versuchsgruppen kann bezüglich ihrer Leistungen im Posttest noch „zu Gute“ gehalten werden, dass sich der Transfertest in dieser Untersuchung unmittelbar an die Experimentalphase anschließt. Ermüdungseffekte wirken sich nicht nur generell auf die Leistung aus, sondern könnten prinzipiell auch Feedbackeffekte untergraben. Allerdings ist die These eines verdeckten Feedbackeffekts unter Berücksichtigung aller bisherigen Ergebnisse und derer Interpretationen doch sehr unwahrscheinlich.

Leistung im Follow-up

Auch im verzögerten Posttest zeigten sich keine Unterschiede in der Leistung der Versuchsgruppen im Lesekompetenztest. Dieses Ergebnis ist vor dem Hintergrund der Interpretationen der Befunde aus dem Experiment und dem unmittelbaren Posttest nicht mehr unplausibel, obwohl in der a-priori aufgestellten Hypothese ein positiver Effekt vermutet wurde. Wenn von den Feedbackinterventionen nicht oder nicht in dem Ausmaß gelernt wird, dass es in der Performanz sichtbar wird, auch nicht im unmittelbaren Posttest, dann wird sich vermutlich auch in einem verzögerten Posttest keine Wirkung zeigen können.

Darüber hinaus ist es für die Zusammenhänge, wie sie sich für dieses Experiment letztlich offenbar haben, auch zu hinterfragen, ob das Ausmaß der Zeitverzögerung des Follow-ups mit durchschnittlich vier Wochen nicht zu lang war. Dieses Design ist vor dem Hintergrund der Erwartung gewählt worden, dass sich spätestens im Posttest positive Effekte der elaborierten Feedbackinterventionen nachweisen lassen sollten. Dann wäre das Follow-up ein guter Indikator für die Nachhaltigkeit des Lerneffekts gewesen. Aber in Anbetracht der Schwierigkeiten, wie sie für die Treatmentphase angenommen werden,

und die anscheinend auch die Voraussetzungen für ein Wirken im unmittelbar anschließenden Posttest verschlechtert haben, erweisen sich vier Wochen Zeitintervall vermutlich als zu groß.

Zwischenfazit zu den Interpretationen der Feedbackauswirkungen

Unter Berücksichtigung aller Resultate wird der Schluss gezogen, dass keine der elaborierten Feedbackarten einen Lerneffekt in der Treatmentphase erzielt hat und somit auch keine Leistungssteigerungen in den Posttests möglich waren. Mögliche Gründe für die unerwartet ausgebliebenen Lerneffekte wurden insbesondere im Zusammenhang mit den Erstantworten der Treatmentphase bereits angerissen. Diese Überlegungen ziehen im Wesentlichen zwei Faktoren als Ursachen für die Wirkungslosigkeit der Feedbacks in Betracht: die Unbrauchbarkeit der Inhalte der drei elaborierten Feedbackarten einerseits und die fehlende Bereitschaft/Motivation der Leser zur Umsetzung der Rückmeldungen andererseits. Zudem werden Merkmale des Untersuchungskontextes als mögliche nachteilige Einflüsse auf die Lernwirkung in Betracht gezogen; die Kontextmerkmale sind aber insofern als Erklärung für die ausgebliebenen Effekte unspezifisch, weil sie sowohl die Nutzbarkeit (Faktor Inhalte) als auch die Nutzung (Faktor Motivation) der Feedbacks untergraben haben könnten. Die Überlegungen zu den möglichen Ursachen für die nicht eingetretenen Lerneffekte werden im Folgenden diskutiert.

7.2 Übergreifende Erklärungen für die Wirkungslosigkeit der elaborierten Feedbackinterventionen

Als mögliche Ursache für die nicht eingetretenen Lerneffekte der elaborierten Feedbackarten werden im Wesentlichen zwei (konkurrierende) Hypothesen in Betracht gezogen: die Feedbackinhalte sind unbrauchbar oder die Motivation der Lerner, diese Inhalte umzusetzen, ist unzureichend. Es wird also davon ausgegangen, dass die Umsetzung/die Nutzung von Feedback vom puren Inhalt trennbar ist. Zwar führt auch eine unbrauchbare Information dazu, dass diese nicht keine Umsetzung auf die geforderten Prozesse nach sich zieht. Aber auch das „nützlichste“ Feedback kann vom Lerner verweigert werden.

7.2.1 Fehlende Nützlichkeit der Feedbackinhalte

Ein Feedback ist dann inhaltlich unbrauchbar, wenn es für die Prozesse, die es stimulieren soll, unpassende, falsche oder unzureichende Informationen bietet. Obwohl die hier untersuchten elaborierten Feedbacks vor dem Hintergrund der Forschung zu Feedback und dem Textverstehen ausgewählt wurden, ist nun abzuwägen, inwieweit sie möglicherweise doch unzureichend sind.

Die Einschätzungen zur Nützlichkeit der Rückmeldungen, die die Probanden der Feedbackbedingungen nach dem Experiment vorgenommen hatten, helfen zur Erklärung nur bedingt. Zum einen führten die unterschiedlichen Feedbackbedingungen (auch Knowledge of Result) zu sehr ähnlichen Einschätzungen. Es waren keine bedeutsamen Gruppenunterschiede erkennbar. Zum anderen fielen die Bewertungen mit durchschnittlich etwa 2.80 Punkten auf der vierstufigen Skala im Allgemeinen eher positiv aus. Dieses Ergebnis erscheint aber nicht unvereinbar mit der generellen Befund der Ineffektivität der Feedbackinterventionen: die subjektiv wahrgenommene Nützlichkeit muss nicht mit dem tatsächlichen, objektiven Nutzen übereinstimmen. So wiesen die Einschätzung auch keine Zusammenhänge mit den Leistungen in den Erstantworten, den Zweitantworten oder der Posttestleistung auf.

Die Nützlichkeit der Feedbacks ist in diesem Zusammenhang eher aus theoretischer Perspektive zu erörtern. Der metakognitive Prompt bezieht sich auf die Überwachung des Verständnisses des Gelesenen. Die metakognitive Überwachung ist vor allem auf der Textebene eine zentrale Komponente des erfolgreichen Textverstehens. Die Wirkung des Prompts erfordert, dass der Leser, anregt durch den Hinweis, sein Verständnis bezüglich des Textes hinterfragt und (meta-)kognitive Strategien der Kontrolle und Regulation des Leseprozesses einsetzt. Das setzt voraus, dass der Leser über geeignete Strategien verfügt und diese angemessen ausführen kann (Brown, 1984; Schnotz, 1991). Das ist voraussetzungsvoll. Aber es kann im Allgemeinen davon ausgegangen werden, dass Leser der untersuchten Altersgruppe der 11-/12-Jährigen (6. Klassen) in einem gewissen Ausmaß schon über relevante Strategien zur Überwachung und Regulation des Leseprozesses verfügen (vgl. Artelt et al., 2005). Das spricht für die Nutzbarkeit des metakognitiven Prompts, wie es auch aus Sicht der Forschung zu Prompting (vgl. Abschnitt 3.5.1.1) zu vertreten ist.

Allerdings ist auch zu bedenken, dass eine gewisse Abhängigkeit zwischen Feedbackgabe und defizitären Fähigkeiten der Leser besteht. Das Feedback wurde bei Fehlern gegeben. Die Fehler reflektieren im Allgemeinen Verständnisschwierigkeiten, die sehr wahrscheinlich auf defizitäre (meta-)kognitive Fähigkeiten in der Textverarbeitung zurückgehen (vgl. Abschnitt 2.3). Damit liegt der Anknüpfungspunkt des metakognitiven Prompts wahrscheinlich an defizitären (meta-)kognitiven Fähigkeiten, die gleichzeitig aber unzureichende (meta-)kognitive Voraussetzungen für die Umsetzung des Prompts darstellen. Damit ist nicht gemeint, dass metakognitive Hinweise ungeeignet sind, um das Textverstehen zu unterstützen. Aber möglicherweise sind sie in dieser Form ohne weitere Hilfestellungen bzw. Instruktionsmaßnahmen (vgl. auch Bannert, 2009) und in der verhältnismäßig kurzen Intervention nicht ausreichend.

Diese Einschätzung kann auch auf das Feedback Fehlererklärung übertragen werden. Die Fehlererklärung erfordert vom Leser ebenso den selbstregulierten Einsatz kognitiver Suchprozesse und metakognitiver Strategien, um darüber zur Lösung zu gelangen. Mit Blick auf eine geforderte Transferleistung muss der Rezipient darüber hinaus aus den Eigenschaften der gemachten Fehler Prinzipien ableiten (z.B. beim Lesen darauf zu achten, Ursache und Wirkung nicht zu vertauschen, genaues (Nach-)Lesen) und auf entsprechende Aufgaben übertragen. Die kognitiven Anforderungen an den Leser und die Voraussetzungen an sein metakognitives Wissen und Können sind ähnlich gelagert wie beim metakognitiven Prompt. Es spricht einiges dafür, dass die Fehlererklärung allein keine hinreichende Hilfestellung ist, sondern zumindest noch weitere Hilfestellungen und/oder Instruktionsmaßnahmen benötigt. Der Blick in die Feedbackliteratur hilft hier allerdings nicht weiter, denn diese Form der Rückmeldung scheint bisher nur in Kombination mit anderen Rückmeldungen umgesetzt (Narciss & Huth, 2004; vgl. auch Narciss, Koerndle & Proske, 2006). Inwiefern die Fehlererklärung allein effektiv ist, ist nicht bekannt.

Das dritte elaborierte Feedback, die Inferenzprompts, bieten die spezifischen Informationen, die in den beiden anderen per se nicht enthalten sind. Sie informieren darüber, wie die Informationen im Text miteinander oder mit dem Hintergrundwissen zu verbinden sind, um den konkret erfragten Zusammenhang herzustellen. Dass sich die Inferenzprompts ebenfalls als nicht effektiv erwiesen haben, lässt sich nicht gut mit der Feedbackliteratur vereinbaren. Zwar erfordern die Inferenzprompts für den Transfer auch

eine gewisse Abstraktionsleistung. Denn dafür müssen die Rückmeldungen, die spezifisch auf die jeweilige Aufgabe angepasst sind, vom konkreten Text bzw. Aufgabeninhalt gelöst werden, um so den zugrunde liegenden Prozess zu erkennen. Aber selbst wenn die Verarbeitung der Prompts nicht so weit erfolgte und der Transfer (deshalb) ausblieb, hätte sich die Wirkung des Feedbacks in den zweiten Antwortversuchen als Leistungsvorteil gegenüber dem einfachen Feedback Knowledge of Result zeigen sollen.

Kognitive Hinweise wie die Inferenzprompts gelten in der Feedbackliteratur als eine sehr effektive Rückmeldung und werden daher meist empfohlen (Bangert-Drowns et al., 1991; Hattie & Timperley, 2007, vgl. auch Abschnit 3.5.1.1). Im Bereich des Textverstehens ist sie mit ähnlicher inhaltlicher Ausrichtung bei Winne und Kollegen (1993) erfolgreich umgesetzt. Allerdings handelte es sich bei dieser Untersuchung um eine länger angelegte Intervention und die Rückmeldungen bestand nicht nur in der Mitteilung, wie die Inferenz zu ziehen ist, sondern der Testleiter (Tutor) markierte die relevanten Informationen im Text und zeigte dem Leser vor allem auch, wie die Inferenz zu ziehen ist.

Insofern ist für die Inferenzprompts, ähnlich wie bei den beiden anderen elaborierten Feedbackarten, zu hinterfragen, ob sie mit ihren Hilfestellungen zu kurz greifen. Es ist sicherlich unstrittig, dass beispielsweise die Kombination mit einem Lesestrategietraining oder die Nutzung der Prompts im Rahmen des Tutoring verspricht gewinnbringend für das Textverstehen/die Lesekompetenz zu sein (vgl. Abschnitt 3.5.1.1). Für den Transfer der Inferenzprompts erscheint das auch nachvollziehbar, für die Leistung in den zweiten Antwortversuchen dagegen nur bedingt.

Die Leistung in den Zweitversuchen hängt davon ab, wie gut die Leser die Rückmeldungen zur Korrektur der Aufgabe nutzen (können). Die Inferenzprompts erfordern hier die Ausführung der kognitiven Schritte, die in den Rückmeldungen enthalten sind. Es ist natürlich nicht ausgeschlossen, dass bei einem falschen oder fehlenden Verständnis hinsichtlich (eines Aspektes) eines Sachverhaltes auch diese Handlungsempfehlungen nicht helfen, vor allem wenn die Fähigkeiten zur Textverarbeitung insgesamt schlecht ausgebildet sind. Aber die Stichprobe reflektiert einen Querschnitt aus Lesern der sechsten Klassen, es gab keinen Fokus auf schwache Leser. Insofern ist davon auszugehen, dass die meisten Probanden prinzipiell in der Lage sind, Inferenzen zu ziehen. Deshalb kann auch vermutet werden, dass auf der Grundlage der Inferenzprompts die Sinnzusammenhänge zumindest für die Aufgabe, für die es das

Feedback gab, im Allgemeinen herzustellen sind. Dafür sprechen auch die Erfahrungen aus den Kognitiven Interviews im Vorfeld des Experiments (vgl. Abschnitt 5.4.2).

Für alle drei elaborierten Feedbackarten können Argumente gefunden werden, warum deren Inhalte (möglichweise) nicht ausreichend sind, um das Textverständnis/die Lesekompetenz in einem Setting wie dem der vorliegenden Arbeit zu verbessern. Dabei erscheinen die Hinweise auf eine fehlende Nützlichkeit aus theoretischer Sicht eher in Bezug auf die Feedbacks Fehlererklärung und metakognitiver Prompt plausibel. Dass die Ineffektivität der Inferenzprompts auf deren fehlende Nützlichkeit zurückführbar ist, kann, zumindest die Leistung in den zweiten Antwortversuchen betreffend, nicht vertreten werden.

7.2.2 Ausgebliebene Feedbackumsetzung

Das Ausbleiben der erwarteten Effekte der elaborierten Feedbackarten kann anstatt auf unbrauchbare Inhalte auch auf eine unzureichende Umsetzung der Rückmeldungen zurückgeführt werden. Für eine Wirkung von Feedback ist nicht nur sein Inhalt maßgebend, sondern, wie in Abschnitt 3.5.2 besprochen, auch die Bereitschaft zur Umsetzung der Informationen durch den Lerner (vgl. auch Bangert-Drowns et al., 1991). Das ist insofern nicht trivial, da die untersuchten Arten elaborierten Feedbacks, wie im vorherigen Abschnitt noch einmal erläutert wurde, eher hohe kognitive Anforderungen stellen.

Zur Erklärung der Bereitschaft/Motivation zur Umsetzung von Feedback kann auf die (erweiterten) Erwartungs-Wert-Modelle der Motivationspsychologie (vgl. Beckmann, J. & Heckhausen, 2006) zurückgegriffen werden. Verkürzt lässt sich festhalten, dass die Motivation von der Wertkomponente (d.h. Vergnügen, Wichtigkeit, Nützlichkeit) und der Erwartungskomponente (d.h. subjektive Wahrscheinlichkeit der gelingenden Umsetzung) abhängt. Die Wert- und Erwartungskognitionen wiederum werden durch motivationale Überzeugungen wie Zielorientierung, Selbstkonzept oder Selbstwirksamkeit beeinflusst. Dabei ist die Motivation zur Feedbackrezeption auch nicht losgelöst von der Lesemotivation (vgl. Möller & Schiefele, 2004) zu betrachten.

Inwieweit die Motivation der Probanden zur Verarbeitung der Rückmeldungen bestanden hat, kann anhand der zur Verfügung stehenden Daten nicht überprüft werden. Auch die Bearbeitungszeiten, die zumindest Hinweise auf die Persistenz der Auseinandersetzung

mit den Rückmeldungen oder den Aufgaben (im Vergleich zu den Kontrollbedingungen) liefern könnten, sind hier nicht dienlich. Zwischen den Bearbeitungszeiten der Versuchsgruppen waren keine statistisch bedeutsamen Unterschiede feststellbar.

Allerdings dürfte der Untersuchungskontext der Motivation zur Feedbackumsetzung zumindest nicht zuträglich gewesen sein. Das Experiment ist in gewisser Weise eher dekontextualisiert, der Nutzen der Verarbeitung der Rückmeldungen ist für Leser nicht unbedingt ersichtlich und die Bearbeitung des Tests hatte für die Probanden auch keine Konsequenzen (z.B. Noten). Zudem ist zu erwarten, dass die eher komplexen kognitiven Anforderungen des Tests und der Feedbacks der Motivation der Probanden im Allgemeinen eher abträglich waren (Clark, Howard & Early, 2006).

Ein weiterer Anhaltspunkt dafür, dass die Wirkungslosigkeit der elaborierten Feedbackarten an der fehlenden Umsetzung dieser liegt, wird in den im Vorfeld der Untersuchung durchgeführten Kognitiven Interviews gesehen. Die elaborierten Feedbacks, insbesondere die Inferenzprompts, zeichneten sich da als nützliche Interventionen für eine Verbesserung des Textverständnisses ab. Hilfreich in Bezug auf die Interpretation der Resultate des vorliegenden Experiments sind Vergleiche zwischen beiden Untersuchungssituationen. Das Setting der Kognitiven Interviews hebt sich dadurch vom Experiment ab, dass die Interviews immer in Einzelsitzungen (ein Testleiter, ein Schüler) und vollständig papierbasiert durchgeführt worden waren. Der Testleiter saß neben dem Schüler und war für die Feedbackgabe durch das Vorlegen entsprechend vorbereiteter Karten zuständig. Außerdem fand ein Austausch über verschiedene Aspekte des Materials statt. Aus der Perspektive des Schülers unterschied sich die Bearbeitungssituation damit im Wesentlich dadurch, dass seine Testbearbeitung unter Beobachtung stand und durch die persönliche Feedbackübermittlung und die Rückfragen des Testleiters ein soziales Setting aufgebaut war. Im Experiment war zwar auch ein Testleiter in der Gruppe anwesend, aber die Probanden arbeiteten für sich, quasi „anonym“, am Programm und das Feedback wird auch „anonym“ über den Computer vermittelt. Damit mag das Setting des Experiments eher wenig unterstützend oder anregend gewirkt haben, um die hohen kognitiven Anforderungen der Feedbackinterventionen in Bezug auf die ebenfalls anspruchsvollen Anforderungen des Textverstehens entsprechend motiviert anzugehen. Das Setting könnte vielmehr noch ein „Ausweichen“ der Probanden vor der Umsetzung der Feedbackinterventionen unterstützt haben. Denn im Verlauf des computerbasierten Experiments gab es keine Kontrolle, ob und inwiefern die Rückmeldungen versucht wurden umzusetzen.

7.3 Diskussion der Untersuchungsmethodik

Bei den eingesetzten Lesekompetenzitems handelt es sich um Eigenkonstruktionen, die im Vorfeld der Untersuchung nicht pilotiert wurden, so wie es in den Vorbereitungen von Kompetenzerhebungen normalerweise der Fall ist. Die Materialien der Untersuchung wurden dagegen in Kognitiven Interviews hinsichtlich verschiedener Aspekte (Lösungshäufigkeiten, Verständlichkeit, Angemessenheit, etc.) überprüft. Dieses Vorgehen ist deshalb vertretbar, weil für das Experiment – als Grundlage für die Feedbackinterventionen – keine Items mit bestmöglichen Testkennwerten notwendig sind. Wichtig für die Intervention ist zum einen, dass die Items generell nicht zu leicht sind, weil dann nur sehr wenige Feedbackgaben und damit Lerngelegenheiten möglich gewesen wären. Zum anderen sollten die Items Eindimensionalität widerspiegeln, damit die Testleistungen als Summenwerte zusammenfassbar sind. Aus inhaltlicher Sicht hätte eine Mehrdimensionalität des Tests Fragen bezüglich der Passung der elaborierten Feedbackart der Inferenzprompts aufgeworfen.
Zudem konnten auch die Testleistungen der feedbackfreien Kontrollgruppe genutzt werden, um die „Qualität“ der Lesekompetenzitems zu bestimmen und gänzlich unpassende Items aus den Analysen auszuschließen. Insgesamt wurden vier der 51 Items aus Experiment und Posttest ausgeschlossen. Des Weiteren bestätigten die Analysen die Eindimensionalität der Lesekompetenztests zu den verschiedenen Messzeitpunkten. Einschränkungen für die Analysen und/oder Interpretationen der Ergebnisse sind aus dem Vorgehen bei der Itementwicklung also nicht anzunehmen.

Bezüglich der Lesekompetenzitems ist ein weiterer Aspekt hervorzuheben: bei der Aufgabenkonstruktion wurde der Schwerpunkt auf inferenzielle Prozesse gelegt, weil sie einerseits ein integraler Bestandteil erfolgreichen Textverstehens sind. Andererseits ermöglicht diese Gruppe von Teilprozessen, dass trotz Variation in den Aufgabenstellungen im Kern eher ähnliche Anforderungen formuliert werden können. Dadurch ist es wiederum möglich, die Feedbackinhalte pro Treatmentgruppe möglichst homogen zu gestalten.

Diskussionswürdig ist auch die Bereinigung der Auswertungsstichprobe(n) um die so genannte „Durchklicker“: Fälle, in denen mehr als 14 % der Items des Treatments oder des Posttest in weniger als sechs Sekunden beantwortet wurden. Dabei kann sowohl der

Cut-off-Wert von fünf Sekunden (5 Sekunden und weniger = durchgeklickt) als auch die Menge der Items, für die ein zu schnelles Antworten erlaubt sein soll, kritisiert werden. Die Entscheidung für diese und gegen andere Kriterien wurde an entsprechender Stelle (vgl. Abschnitt 5.8.1) dargelegt. Aus den dazu angeführten Häufigkeitsverteilungen wird auch ersichtlich, dass liberalere oder strengere Kriterien nicht zu extrem anderen Fallausschlüssen führen würden.
Auffällig ist jedoch der hohe Prozentsatz der nach dem gewählten Kriterium ausgeschlossenen Fälle. Für die Analysen des Treatments entfallen ungefähr 12 % der Stichprobe und für den Posttest sogar ein Drittel der Probanden. Daraus ist nicht die Schlussfolgerung zu ziehen, dass die Ausschlusskriterien liberaler gestaltet oder Probanden generell nicht aufgrund ihrer Bearbeitungszeiten ausgeschlossen werden sollten. Denn zum einen ist der gewählte Cut-off-Wert von fünf Sekunden schon ein Minimalkriterium; es basiert auf Erfahrungswerten (Ausprobieren). Zum anderen ist das zu schnelle Antworten, das eine hinreichend ernsthafte Testbearbeitung bezweifeln lässt, ein Problem, das vorhanden ist. Die Feedbackinterventionen können nur wirken, wenn die Probanden sie mit den Rückmeldungen auseinandersetzen. Das Setting der Untersuchung mag hat diese Verhalten vielleicht begünstigt. Andererseits ist es auch möglich, dass es sich hier um ein generelles Problem im Zusammenhang mit der Bearbeitung von (Kompetenz-)Tests handelt, das durch die computerbasierte Administrierung und die Erfassung der Bearbeitungszeiten in dieser Untersuchung offensichtlich wird.

Als letzter Aspekt der Methodik der Untersuchung wird noch auf die Repräsentativität der Stichprobe eingegangen. Diese ist insofern eingeschränkt, da zunächst die für eine Teilnahme angefragten Schulen und dann jeder Schüler bzw. seine Erziehungsberechtigten frei über die Teilnahme entschieden haben. Positiv hervorzuheben ist, dass die gängigen Schulformen alle berücksichtigt wurden.

7.4 Gesamtfazit und Ausblick auf das zweite Experiment

Die elaborierten Feedbackarten – Fehlererklärung, metakognitiver Prompt und Inferenzprompts – sind entgegen den Erwartungen wirkungslos geblieben. Sie konnten keinen Lerneffekt nach sich ziehen, der gegenüber dem einfachen Feedback Knowledge of Result und der feedbackfreien Kontrollbedingung eine Leistungssteigerung

hervorgerufen hätte. Das einfache Feedback Knowledge of Result zeigte sich ebenfalls nicht nützlich, um die Testleistung anzuheben, und bestätigte damit die Erwartungen.

Zur Erklärung der Ineffektivität der elaborierten Feedbackarten wurden verschiedene Gründe in Betracht gezogen. Sie lassen sich im Wesentlichen in zwei (konkurrierenden) Hypothesen zusammenfassen: die Feedbackinhalte waren unbrauchbar oder die Motivation der Lerner, diese Inhalte umzusetzen, war unzureichend. Für beide Erklärungsansätze konnten Argumente gefunden werden. Auf der einen Seite ist offenkundig, dass die Wirkung vor allem der elaborierten Feedbacks Fehlererklärung und metakognitiver Prompt hohe kognitive Anforderungen an die Lerner stellt. Sie erfordern in weiten Teilen den selbstregulierten Einsatz angemessener (meta-)kognitiver Strategien, was sich insbesondere auch vor dem Hintergrund des Anknüpfens von Feedback an Verständnisschwierigkeiten als zu voraussetzungsvoll herausgestellt haben könnte. Bezüglich der Inferenzprompts wurde die Lage anders eingeschätzt. Dieses Feedback bietet die kognitiven Hinweise, die zum Herstellen von Sinnzusammenhängen erforderlich sind; hier konnten keine plausiblen Argumente gefunden werden, warum der Inhalt der Inferenzprompts unzureichend oder ungeeignet sein sollte, um nicht wenigstens der unmittelbaren Korrektur falscher Erstantworten dienlich zu sein.
Auf der anderen Seite spricht insbesondere der Charakter der Untersuchungssituation dafür, dass die Motivation der Lerner zur Umsetzung der Rückmeldungen eher untergraben worden sein könnte. Als nachteilig erwies sich hier sicherlich, dass den Schülern ein Nutzen der Feedbackverarbeitung nicht erkennbar bzw. nicht gegeben war. Auch die Erwartungen bezüglich der eigenen Leistungsverbesserungen nach Feedback mögen aufgrund der kognitiv eher anspruchsvollen Anforderungen bei vielen Lesern nicht hoch ausgefallen sein. Motivationale Effekte zu Ungunsten der Feedbackverarbeitung sind in der Konsequenz also auch nicht losgelöst von den zugrunde liegenden kognitiven Anforderungen des erfassten Fähigkeitsbereiches und den Feedbackinhalten zu betrachten. Hinweise, dass sich die Feedbackinterventionen nachteilig auf die Aufgeregtheit oder Besorgtheit (Testangst) der Probanden ausgewirkt hätte, gibt es keine.

Letztlich bleiben alle Überlegungen zu den Gründen der Ineffektivität der elaborierten Feedbackarten aber Spekulationen, denn eine Überprüfung ist anhand der zur Verfügung stehenden Untersuchungsdaten nicht möglich. Auch die Feedbackliteratur bietet hier

kaum differenzierte Anhaltspunkte, geschweige denn (differenzierte) Modelle bzw. Theorien zur Wirkung und eben Wirkungslosigkeit von Feedbackinterventionen. Aber unter Berücksichtigung aller Resultate und Untersuchungsmerkmale wird dem Faktor der fehlenden/unzureichenden Motivation zur Umsetzung der Rückmeldungen, zumindest in Bezug auf die Inferenzprompts, ein größeres Gewicht beigemessen. Entscheidend für diese Einschätzung sind die Erfahrungen der Kognitiven Interviews, die positive Effekte für die elaborierten Feedbacks, insbesondere die Inferenzprompts, vermuten ließen. Das Setting der Kognitiven Interviews hebt sich gegenüber dem Experiment in einigen Punkten ab, denen vor dem Hintergrund der erfolglosen Feedbackinterventionen im Experiment im Nachhinein Aufmerksamkeit geschenkt wird.

Der wesentliche Unterschied ist, dass in den Kognitiven Interviews das Feedback durch den Testleiter in einem Eins-zu-Eins-Setting gegeben wurde. Der Einsatz des Testleiters als Feedbackgeber war für die Interviews eine primär pragmatische Entscheidung. Nun stellt es sich jedoch so dar, dass ein Testleiter als Feedbackgeber bzw. das damit verbundene soziale Setting einen günstigen Rahmen für die Umsetzung der Feedbackinterventionen seitens der Probanden geschaffen haben, wodurch die Rückmeldungen dann wiederum ihr Potential in der Unterstützung des Textverständnisses/der Lesekompetenz entfalten konnten. Natürlich kann nicht davon ausgegangen werden, dass die Übertragung des Testleiters als Feedbackgeber in den Kontext des Experiments die Wirkung der elaborierten Feedbacks garantiert. Denn der Umfang der Materialien im Experiment und der Testcharakter heben sich ebenso von den Kognitiven Interviews ab. Aber unter Berücksichtigung aller Umstände besteht die begründete Annahme, dass eine Änderung des Testsettings Feedbackeffekte hervorbringen kann. Um diese Annahme zu überprüfen (und die Ergebnisse des vorliegenden Experiments gegebenenfalls abrunden zu können) wird ein zweites Experiment durchgeführt, das in weiten Teilen dem vorliegenden Experiment gleicht, zusätzlich aber eine Feedbackintervention realisiert, in der die Lerner das Feedback von einer Person erhalten.

Experiment 2 (Fokus: Kontrastierung des Präsentationsmodus)

Die Rolle des vorliegenden Experiments im Rahmen dieser Arbeit besteht darin, der Ineffektivität der Feedbackinterventionen bzw. des Untersuchungskontextes im ersten Experiment nachzugehen. Die Diskussion möglicher Gründe für das Ausbleiben der Effekte rückte einen Faktor in den Vordergrund: eine zu geringe Motivation bzw. eine durch den Untersuchungskontext begünstigte, unzureichende Umsetzung der Rückmeldungen. Dieser Aspekt wurde im vorliegenden Experiment aufgegriffen, indem die Bedingungen der Feedbackgabe bzw. der Untersuchungskontext abgewandelt wurden.

Während im ersten Experiment bewusst eine ausschließlich über den Computer vermittelte Feedbackgabe umgesetzt wurde (vgl. Abschnitt 3.5.1.3), wird diese im vorliegenden Experiment mit der Feedbackpräsentation über eine Person, den jeweiligen Testleiter, kontrastiert. Diese neue Variante der Feedbackpräsentation soll die Umsetzung der Feedbackmitteilungen gewährleisten, die bei der rein computerbasierten Intervention möglicherweise unterwandert wird und so die ausbleibenden Effekte im ersten Experiment bewirkt haben könnte (vgl. Abschnitt 7). Dass dabei auf den Testleiter als Feedbackgeber zurückgegriffen wird, ist aus den Erfahrungen der noch vor dem ersten Experiment durchgeführten kognitiven Interviews abgeleitet. Vor dem Hintergrund der bestehenden Literatur (vgl. Abschnitt 3.5.1.3) ist diese Art der Feedbackpräsentation allerdings auch nicht unkritisch. Dieser Punkt sowie einige weitere einführende Bemerkungen zum vorliegenden Experiment werden vorab in Hinblick auf die Einführung der Fragestellungen und Hypothesen erläutert.

8 Vorbemerkungen

Der Testleiter als Feedbackgeber

Der neue Präsentationsmodus, um den das zweite Experiment ergänzt wird, sieht also vor, dass die Rückmeldungen in der Treatmentphase persönlich über den Testleiter, nicht über den Computer gegeben werden. Die Intervention durch den Testleiter beschränkt sich dabei ausschließlich auf das Geben der festgelegten Rückmeldungen, die in ihrem Wortlaut identisch mit den rein computervermittelten Feedbacks sind. Das impliziert

auch, dass kein Dialog zwischen Testleiter und Proband stattfindet und so beispielsweise auch keine Rückfragen zum Verständnis eines erhaltenen Feedbacks vorgesehen sind. Die Feedbackgabe durch den Testleiter erfolgt in einer Kombination verbaler und schriftlicher Darstellung. Für die entsprechende Versuchsgruppe zieht dies aus praktischen Gründen (etwa gegenseitige Störungen durch das Sprechen der Testleiter) gleichwohl die Durchführung des Experiments in Einzelsitzungen nach sich. Über diese Aspekte hinaus ist die Bedingung der testleitergebundenen Feedbackgabe nicht von dem bisherigen computervermittelten Präsentationsmodus verschieden (für die ausführliche Methodik vgl. Abschnitte 10.2 und 10.7.2).

Der Ansatz, den Testleiter das Feedback geben zu lassen, ist an die Erwartung geknüpft, dass damit der anspruchsvolle Prozess der Feedbackverarbeitung gestützt wird. Durch die persönliche Feedbackgabe erhält der Proband ein soziales „Gegenüber", das Feedback sollte dadurch weniger anonym vermittelt wahrgenommen werden. Vor allem kann dadurch aber eine Testsituation entstehen, in der der Feedbackverarbeitung eine höhere Geltung oder Bedeutsamkeit zugeschrieben wird (Stichwort *Accountability*). Der Proband kann hier einen direkten Bezug zwischen den Rückmeldungen und einer Person herstellen, die für diese Hilfestellungen steht und diese dem Anschein nach sogar passend für die Problemstellungen des Lerners formuliert/erdacht haben mag. Die Anstrengungsbereitschaft zur Umsetzung der Feedbackmitteilungen sollte in solch einem Setting höher ausfallen als bei der rein computervermittelten Feedbackgabe. Darüber hinaus fällt es Probanden mit einem unmittelbar anwesenden Testleiter ungleich schwerer, Rückmeldungen etwa gänzlich zu ignorieren und „weiterzuklicken". Letztlich hebt sich die Interventionssituation mit einem Testleiter als Feedbackgeber von der rein computervermittelten Feedbackintervention durch ein Konglomerat an Merkmalen ab, die nicht voneinander trennbar sind. Diesen Umstand gilt es bei der Interpretation und Diskussion der Effekte der Intervention zu berücksichtigen.

Vor dem Hintergrund der im Abschnitt 3.5.1.3 dargelegten Forschungsarbeiten ist der Einsatz des Testleiters als Feedbackgeber jedoch nicht unkritisch: vor allem durch das persönliche Gegenüber in der Intervention besteht auf Seiten des Feedbackempfängers das Risiko gewisser Abwehrtendenzen oder aufgabenirrelevanter Kognitionen, die sich leistungsmindernd auswirken können (Ashford & Cummings, 1983; Kluger & DeNisi, 1996). Das Feedback wird dann stärker als Bewertung der eigenen Person bzw. der eigenen Leistungsfähigkeit wahrgenommen und das löst beim Lerner eher Besorgnis aus (Kluger & Adler, 1993) und/oder lenkt seine Aufmerksamkeit auf meta-tasks (Kluger &

DeNisi, 1996). Damit ist gemeint, dass der Fokus der Aufmerksamkeit auf das Ich oder auf affektive Prozesse gelenkt wird und damit nicht mehr für die Bewältigung der vorliegenden Aufgaben zur Verfügung steht. Diese Mechanismen greifen vermutlich umso mehr bei negativem, das heißt fehlerbezogenem Feedback (Comer, 2007). Auch die Studie von Kluger und Adler (1993) spricht dafür, dass ein Testleiter, der nicht nur als Beobachter, sondern als Feedbackgeber fungiert, leistungsmindernde Effekte nach sich zieht. Gleichzeitig liefert die Studie von Kluger und Adler aber wiederum auch beschwichtigende Argumente, denn mit einem Testleiter als Feedbackgeber fällt die Leistung zumindest nicht schlechter aus, als wenn das Feedback ausschließlich über den Computer vermittelt wird und kein Beobachter anwesend ist. Zudem werden die Erwartungen an die testleitervermittelte Feedbackgabe durch die für den Bereich der Lesekompetenz dargelegten Studien gestärkt, die ihre Feedbackinterventionen ebenfalls über eine Person vermitteln ließen und positive Effekte auf die Leistung aufweisen. Das bedeutet, obwohl bei der testleitergebundenen Feedbackgabe ambivalente Wahrnehmungen oder Reaktionen auf Seiten des Lerners nicht auszuschließen sind, können für die Intervention dennoch positive Effekte auf die Leistung erwartet werden.

Das Feedback

Im vorliegenden Experiment wird auf die Feedbackarten des ersten Experiments zurückgegriffen, wobei aus ökonomischen Gründen nur zwei der vier Feedbackarten ausgewählt werden. Zum einen wird die elaborierte Feedbackart Inferenzprompt wieder aufgegriffen, da hier vor dem theoretischen Hintergrund das größte Potenzial im Vergleich zu den anderen elaborierten Feedbackarten, wie sie im ersten Experiment eingesetzt worden sind, vermutet wird (vgl. Abschnitt 3.5.1.1). Zum anderen wird auch die einfachste Feedbackart Knowledge of Result wiederholt, da sie als Rückmeldung mit geringstmöglichem Informationsgehalt, aber äquivalenter Bearbeitungsprozedur wie die Intervention mit elaboriertem Feedback den besten Kontrast zur Bestimmung des Mehrwerts der elaborierten Informationskomponente darstellt. Als weiterer Kontrast wird wiederum eine Kontrollbedingung ohne jegliche Intervention geprüft, die die für bestimmte Auswertungs- bzw. Interpretationsoptionen relevante Leistungsfähigkeit der Stichprobe unter konventionellen Testbedingungen abbildet.

Synthese der Präsentationsmodi und Feedbackarten

Die Bedeutung der Feedbackgabe über den Testleiter ergibt sich nicht gleichermaßen für beide Feedbackarten. Die Botschaft bei Knowledge of Result („Das ist falsch.“) ist denkbar knapp und zudem immer dieselbe. Es spricht wenig dafür, dass hier eine unzureichende Rezeption als Ursache für eine Ineffektivität der Intervention problematisch sein könnte, oder vice versa, dass die hohen Anforderungen an eine gewinnbringende Nutzung des Feedbacks durch die Maßnahme der testleitervermittelten Feedbackpräsentation gemeistert werden.

Im Gegensatz dazu enthält die elaborierte Feedbackart, die Inferenzprompts, per Definition vergleichsweise umfangreichere Mitteilungen, die darüber hinaus spezifisch für die jeweilige Aufgabe bzw. den Text formuliert und damit im Wortlaut nie identisch sind. Hier werden vor dem Hintergrund der erläuterten Absicht des Experiments die Notwendigkeit und das Potential der testleitergebundenen Feedbackgabe gesehen. Zusätzlich werden die Inferenzprompts aber auch als rein computervermittelte Intervention dargeboten, obwohl sich diese Bedingung als ineffektiv erwiesen hatte. Da die Untersuchung aber an einer neuen Stichprobe durchgeführt wird und eine unmittelbare Vergleichbarkeit zwischen beiden Experimenten auch deshalb eingeschränkt ist (siehe unten), wird diese Bedingung zugunsten der Aussagekraft des vorliegenden Experiments wiederholt.

Demzufolge besteht das Konzept des vorliegenden Experiments darin, die elaborierte Feedbackart mit beiden Präsentationmodi zu kombinieren: zum einen mit dem Testleiter als Feedbackgeber (nachfolgend kurz: Inferenzprompts-via-Testleiter) und zum anderen mit dem Computer als Feedbackquelle (Inferenzprompts). Die einfache Feedbackart Knowledge of Result wird ausschließlich in der computervermittelten Bedingung dargeboten (Knowledge of Result) und die Kontrollbedingung (kein Feedback) ist analog dazu wieder als selbstständige Arbeit am Computer im Rahmen der Gruppentestung gestaltet.

Vergleichbarkeit zum vorherigen Experiment

Neben dem neu hinzugefügten Präsentationsmodus sowie der erläuterten Auswahl der Feedbackarten unterscheidet sich das vorliegende Experiment in einigen weiteren Punkten vom vorherigen. Die Änderungen sind vornehmlich aus ökonomischen Gründen

getroffen worden und betreffen: 1) die Reduzierung der Stichprobe auf eine Schulform, und zwar die Realschule, 2) die Begrenzung des Untersuchungsumfangs auf eine Untersuchungssitzung, in der das computergestützte Experiment und der Posttest durchgeführt werden, 3) eine Reduzierung der Instrumente zur Erfassung kognitiver oder motivational-emotionaler Konstrukte als Kontrollvariablen und 4) eine Anpassung der Texte und Items. Insbesondere die Wahl der Stichprobe und die Änderungen des Lesematerials schränken die unmittelbare Vergleichbarkeit der Ergebnisse beider Experimente ein.

9 Fragestellungen und Hypothesen

Wie einleitend bereits erwähnt verfolgte das vorliegende Experiment (unter angewandelten Untersuchungsbedingungen) dieselben zentralen Fragestellungen wie das erste Experiment. Die *erste und zentrale Fragestellung* bezieht sich also auf die *Wirksamkeit der Feedbackbedingungen auf die Lesekompetenz/das Textverständnis*. Die Erwartungen bezüglich der Feedbackwirkung spiegeln hinsichtlich der ausschließlich computerbasierten Bedingungen Inferenzprompt und Knowledge of Result die Erfahrungen aus dem ersten Experiment wider: bezüglich der Bedingung Inferenzprompt wird davon ausgegangen, dass diese nicht wirken, weil sie in dem Kontext der Untersuchungssituation im Allgemeinen nicht verarbeitet werden. Hinsichtlich der Bedingung Knowledge of Result wird davon ausgegangen, dass sie vor allem deshalb keinen Effekt auf die Leistung bewirkt, weil der Informationsgehalt von Knowledge of Result für das Textverstehen zu gering und das Feedback deshalb nicht nützlich ist. Der neu hinzugefügten Bedingung Inferenzprompt-via-Testleiter wird aus den oben erläuterten Gründen dagegen das Potential zur Leistungssteigerung mittels Inferenzprompts zugesprochen. Das heißt, es wird angenommen, dass durch die damit verbundene Untersuchungssituation ein Setting geschaffen wird, in dem die Anstrengungsbereitschaft der Probanden verbessert und dadurch die Umsetzung der Rückmeldungen gewährleistet wird, wodurch sich die Effektivität der Inferenzprompts beweisen kann. Die Annahmen lassen sich wie folgt zusammenfassen:

a) Die Probanden in der Bedingung Inferenzprompts-via-Testleiter erzielen eine höhere Leistung als die Probanden aller anderen Bedingungen.
b) Die Leistungen in den Gruppen Inferenzprompts, Knowledge of Result und der feedbackfreien Kontrollgruppe unterscheiden sich hingegen nicht voneinander.

Die formulierten Erwartungen beziehen sich sowohl auf die Leistung in der Treatmentphase (erste und zweite Antwortversuche) als auch den Posttest.

Die *zweite Fragestellung* richtet sich auf die *Bearbeitungszeiten* in der Treatmentphase und im Posttest. Die Bearbeitungszeiten fungieren als grober Indikator für die Annahme der jeweiligen Intervention seitens der Probanden. Damit ist gemeint, dass sich eine

gewisse Auseinandersetzung mit den Rückmeldungen in einer längeren Bearbeitungsdauer des Experiments erkennen lassen sollte. Das erste Experiment hat hierzu gezeigt, dass die ausschließlich computervermittelten Feedbackinterventionen für die Treatmentphase insgesamt nicht mehr Zeit als die Kontrollgruppe aufgebracht hatten. Diese Effekte werden auch für das vorliegende Experiment vermutet. Für die Feedbackbedingung Inferenzprompt-via-Testleiter werden dagegen im Sinne der theoretischen Annahmen längere Bearbeitungszeiten erwartet, und zwar im Treatment (Erst- und Zweitantworten) und im Posttest. Eine längere Bearbeitung der Zweitantworten ist unmittelbar an die Annahme geknüpft, dass die Bedingung zu einer Auseinandersetzung mit den Rückmeldungen beim Beantworten der Zweitantworten führt. Die Umsetzung der Prompts ist als ressourcenfordernder Prozess zu sehen (vgl. Abschnitt 3.5.2), der entsprechend mehr Zeit benötigt. Da davon ausgegangen wird, dass die Informationen der Prompts auch für die Beantwortung der Erstantworten und des Posttests genutzt werden, ist auch hier von einer im Vergleich zu den anderen Bedingungen erhöhten Bearbeitungszeit auszugehen. Die Annahmen lauten also:

a) Die Bedingung Inferenzprompt-via-Testleiter führt zu einer längeren Bearbeitung der Aufgaben als in den anderen Bedingungen.
b) Die Bearbeitungszeiten in den Versuchsgruppen Knowledge of Result und Inferenzprompt sowie der feedbackfreien Kontrollgruppe unterscheiden sich nicht voneinander.

Die *dritte Fragestellung* bezieht sich auf die Auswirkungen der Feedbackinterventionen auf die *Anstrengungsmotivation*. Hier wird ebenfalls vor der Hintergrund der erläuterten Annahmen zu den Auswirkungen der testleitergebundenen Feedbackgabe davon ausgegangen, dass die Bedingung Inferenzprompt-via-Testleiter zu einer höheren Anstrengungsmotivation führt als sie in den anderen Bedingungen (Inferenzprompt, Knowledge of Result und Kontrollgruppe) vorhanden sein wird.

Die *vierte Fragestellung* betrifft schließlich die Auswirkungen der Feedbackinterventionen auf die Einschätzungen der Probanden hinsichtlich der *Nützlichkeit der Rückmeldungen*. Es wird erwartet, dass die Einschätzungen in der Bedingung Inferenzprompt-via- Testleiter positiver ausfallen als die Einschätzungen der beiden computerbasierten Bedingungen, die sich zudem – wie im ersten Experiment – hierin nicht unterscheiden. Der Effekt zugunsten der Testleiterbedingung leitet sich

wiederum aus der Annahme ab, dass in dieser Bedingung die Rezeption und Umsetzung der Feedbacks aufgrund des geschaffenen Settings erst gelingt und sich dann in positiven Effekten auf die Leistung niederschlägt, was von den Probanden auch entsprechend wahrgenommen werden sollte.

10 Methodik

10.1 Stichprobe

An der Untersuchung nahmen insgesamt 251 Schüler der sechsten Klassenstufe teil[7]. Die Stichprobe rekrutierte sich aus neun Realschulen der Regierungsbezirke Ober- und Unterfranken (Bayern). Von den Teilnehmern waren 126 Mädchen (50.2 %) und 125 Jungen (49.8 %). Das durchschnittliche Alter der Probanden betrug 12 Jahre und 5 Monate (*SD* = 0;7 Jahre). Um einen Indikator für den Migrationshintergrund der Probanden zu erhalten, wurden sie gefragt, welche Sprache sie hauptsächlich zu Hause sprechen. Darauf hatten insgesamt 13 Teilnehmer (4.8 %) eine nicht deutsche Sprache angegeben.
Die Teilnahme an der Untersuchung war zunächst den Schulen und dann auch allen Schülern bzw. ihren Erziehungsberechtigten freigestellt worden. Die Untersuchung war im Vorfeld durch das Bayerische Staatsministerium für Unterricht und Kultus genehmigt worden.

10.2 Untersuchungsdesign

Diese Untersuchung beschränkte sich auf die Durchführung des Experiments und eines unmittelbaren Posttests. Ergänzend wurden einige wenige Instrumente zur Erfassung von Hintergrund- bzw. Kontrollvariablen administriert. Auf eine Vortestung und ein Follow-up wurde verzichtet. Damit konnte die Untersuchung in nur einer Sitzung durchgeführt werden.

Das Experiment basierte auf einem zweifaktoriellen, between-subjects Design, von dem nicht alle Stufenkombinationen umgesetzt wurden. Der erste Faktor (A), die „Feedbackart“, war dreistufig gestaltet: Knowledge of Result, Inferenzprompt und kein Feedback. Der zweite Faktor (B), die „Art der Feedbackpräsentation“, war zweistufig

[7] Bedingt durch fehlende Werte und Fallausschlüsse (vgl. Abschnitt 10.8.1) weichen die Auswertungsstich-proben zum Teil deutlich von N = 251 ab.

angelegt: hier wurde die rein computerbasierte Darbietung der Rückmeldungen mit dem Testleiter als Feedbackgeber verglichen. Von den 3*2 möglichen Versuchsbedingungen wurden vier Versuchsgruppen realisiert:

1. *Experimentalbedingung A_1B_1: „Knowledge of Result" via Computer:*

   Das ist die Mitteilung, dass die gewählte Antwort falsch ist („Das ist falsch."). Das Feedback wurde über den Computer vermittelt. Diese experimentelle Bedingung wurde damit exakt wie im ersten Experiment umgesetzt.

2. *Experimentalbedingung A_2B_1: „Inferenzprompt" via Computer:*

   Das ist „Knowledge of Result" kombiniert mit einem Hinweis, wie die durch die Fragestellung geforderte Inferenz gezogen werden kann. Dabei wird aufgefordert, den Zusammenhang zwischen Ereignisse herzustellen, eine zeitliche Abfolge von Ereignissen zu überprüfen, was die Ursache einer Situation ist und/oder worin die Wirkung eines Ereignisses liegt (für Beispiele siehe Abschnitt 5.1). Die konkreten Rückmeldungen sind für alle Items auf die jeweilige Fragestellung ausgerichtet. Die Rückmeldungen wurden über den Computer vermittelt. Damit wurde diese Bedingung so wie im ersten Experiment umgesetzt.

3. *Experimentalbedingung A_2B_2: „Inferenzprompt-via-Testleiter":*

   Der Feedbackinhalt ist derselbe wie in der vorher erläuterten Bedingung Inferenzprompt via Computer. Doch die Rückmeldungen wurden nicht über den Computer, sondern durch den anwesenden Testleiter gegeben.

4. *Kontrollbedingung A_3B_1 :* kein Feedback:

   Es wurde kein Feedback gegeben und die Probanden arbeiteten wie die ersten beiden Versuchsgruppen selbstständig am Computer.

Die Zuteilung der Untersuchungsteilnehmer zu den Versuchsbedingungen erfolgte innerhalb jeder untersuchten Schulklasse randomisiert. Als abhängige Variable wurde die Lesekompetenz im Experiment und im Posttest erfasst.

10.3 Instrumente

Die in der vorliegenden Untersuchung eingesetzten Instrumente wurden im Vergleich zum ersten Experiment deutlich reduziert, da die Untersuchung in einer einzigen Sitzung (Doppelstunde) durchgeführt wurde. Von den Instrumenten des ersten Experiments (vgl. Abschnitt 5.3) wurden der Lesegeschwindigkeitstest, der Fragebogen zur Testangst

sowie die eigenentwickelten Items zur wahrgenommenen Nützlichkeit der Rückmeldungen wiederholt und um ein Instrument zur Erfassung der Anstrengungsmotivation ergänzt.

Lesegeschwindigkeit und Testangst stellen zwei wichtige Einflussfaktoren auf die Leistungsfähigkeit im (feedbackgestützten) Leseverständnistest dar und wurden deshalb wieder als Kontrollvariablen vor der Durchführung des Experiments erfasst (vgl. Abschnitt 5.3 für Begründung des Einsatzes der Instrumente). Somit konnten die Versuchsgruppen, zusätzlich zur durchgeführten Randomisierung, im Vorfeld der Leistungsanalysen auf die Gleichverteilung hinsichtlich dieser beiden Maße geprüft werden. Das Konstrukt der Testangst wurde den anderen, vor allem motivationalen Konstrukten aus dem ersten Experiment vorgezogen, weil es vor allem in Hinblick auf die Bedingung der testleitergebundenen Feedbackgabe und deren mögliche nachteilige Einflüsse auf die Lerner relevant erscheint.

Das Instrument zur Messung der Anstrengungsmotivation (bezogen auf die Treatmentphase) wurde infolge der Ergebnisse des ersten Experiments und der damit verbundenen Diskussion hinsichtlich der Testmotivation der Probanden eingesetzt. Durch die Erfassung der Anstrengungsmotivation als abhängige Variable können die diesbezüglichen Auswirkungen der Feedbackinterventionen untersucht werden. Die Einschätzung der Nützlichkeit der Rückmeldungen wurde wiederum aus einem inhaltlich-pragmatischem Interesse bezüglich der Wahrnehmung der Feedbacks erfragt (z.B. Hinweise auf Verbesserungspotential), aber auch hinsichtlich möglicher Effekte der Feedbackinterventionen untersucht.

Die Lesegeschwindigkeit und die Testangst wurden vor dem Experiment, die Anstrengungsmotivation und die Nützlichkeit der Rückmeldung danach erfasst. Außer dem Lesegeschwindigkeitstest wurden die Instrumente computerbasiert dargeboten und waren dabei in das Programm zur Erfassung der Lesekompetenz bzw. Darbietung des Treatments eingebettet. Die nachfolgenden Beschreibungen der Instrumente enthalten jeweils auch die in der vorliegenden Stichprobe empirisch gefundenen Kennwerte.

10.3.1 Leistungstest

Lesegeschwindigkeit

Zur Erfassung der Lesegeschwindigkeit wurde wiederum das „Salzburger Lesescreening für die Klassen 5 bis 8“ (SLS 5-8; Auer et al., 2005) als Papier-und-Bleistift Verfahren eingesetzt. Für die Beschreibung des Verfahrens wird auf Abschnitt 5.3.1 verwiesen. Im Unterschied zum ersten Experiment wurden diesmal zwei unterschiedliche Testformen (Formen A1 und A2 aus Auer et al., 2005) eingesetzt, um Unterschleif zu unterbinden. Die beiden Testformen unterscheiden sich in der Reihenfolge der Items. Die Relevanz der Erfassung der Lesegeschwindigkeit im Rahmen der Untersuchung ist ebenfalls in Abschnitt 5.3.1 erläutert.

Tabelle 31 Kennwerte des Lesegeschwindigkeitstests

	Lesegeschwindigkeit (theoretisches Max: 70)
N	249 [a]
Mittelwert	40.06
Standardabweichung	6.93
Minimum	19
Maximum	58

Anmerkungen. [a] Die Auswertungsstichprobe reduzierte sich hier um zwei Fälle, die wegen instruktionswidrigen Verhaltens im Lesegeschwindigkeitstest nicht genutzt wurden.

Die Testkennwerte für diese Untersuchung sind in Tabelle 31 zusammengefasst. Die mittlere Lesegeschwindigkeit beträgt M = 40.06 (SD = 6.93) und fällt im Vergleich zur Normstichprobe des Tests für sechste Klassen in Hauptschule/Realschule (M = 33.3, SD = 6.7) etwas höher aus. Ein Reliabilitätsmaß kann für die vorliegende Stichprobe nicht berechnet werden. Aber die im Manual des Salzburger Lesescreenings angegebene Paralleltest-Reliabilität von r = .89 (Auer et al., 2005, S. 9) zeigt, dass es sich um ein sehr reliables Verfahren handelt.

10.3.2 Fragebogen

Testangst

Zur Erfassung der Testangst (Komponenten Aufgeregtheit und Besorgtheit) wurde wiederum das „Prüfungsängstlichkeitsinventar TAI-G“ (Hodapp, 1991) genutzt. Für die Beschreibung des Verfahrens und die Erläuterung der Relevanz der Erhebung von Testangst im Rahmen dieser Arbeit wird auf Abschnitt 5.3.2 verwiesen. Im Unterschied zum ersten Experiment wurden die 8 Items zur Erfassung der Testangst in diesem Experiment nur einmal, und zwar unmittelbar davor eingesetzt. Das Antwortformat war eine vierstufige Skala von 1 für „stimmt gar nicht“ bis 4 für „stimmt völlig“.

Die Auswertung erfolgte wie im ersten Experiment in Anlehnung an Hodapp (1991) für beide Komponenten bzw. Subskalen getrennt. Die entsprechenden Kennwerte sind in Tabelle 32 zusammengefasst. Die interne Konsistenz für die Subskala Besorgtheit ist mit Cronbachs α = .75 hinreichend, die interne Konsistenz der Subskala Aufgeregtheit fällt mit Cronbachs α = .53 hingegen äußerst niedrig aus. Beide Subskalen korrelieren mit r = .47 (p < .001) mittelhoch miteinander.

Tabelle 32 Kennwerte der Subskalen zur Testangst

	Testangst	
	Subskala Aufgeregtheit (3 Items)	Subskala Besorgtheit (5 Items)
N	251	251
Mittelwert	1.66	2.21
Standardabweichung	0.58	0.67
Cronbachs α	.53	.75

Einschätzung zur Nützlichkeit der Feedbacks

Unmittelbar nach dem Experiment wurden wieder die fünf Items zur Einschätzung der subjektiv wahrgenommenen Nützlichkeit der angebotenen Rückmeldungen computerbasiert dargeboten (vgl. Abschnitt 5.3.2). Im Vergleich zum ersten Experiment wurde die Aufgabenstellung konkretisiert und die Items wurden in ihrem Wortlaut leicht verändert. Im ersten Experiment lautete die Aufgabenstellung: „Bitte schätze ein, wie du die Rückmeldungen in den Texten fandest!“. Sie wurde geändert in: „Wie beurteilst du insgesamt die Hinweise, die du nach falschen Antworten auf die Fragen erhalten hast?“.

Die sprachliche Veränderung der Items bestand darin, in allen Items durchgängig nach den „Hinweisen“ zu fragen anstatt mit dem Pronomen „sie“ lediglich einen referentiellen Bezug zur Aufgabenstellung herzustellen (z.B. „Ich fand *die Hinweise* hilfreich.“ statt „Ich fand *sie* hilfreich.“). Die Items lauteten wie folgt:

1. Ich fand die Hinweise hilfreich.
2. Ich fand die Hinweise verwirrend.
3. Die Hinweise haben mich abgelenkt.
4. Die Hinweise haben mir geholfen, die Aufgabe doch noch zu bewältigen.
5. Die Hinweise haben mir bei darauffolgenden Aufgaben geholfen.

Die Beantwortung der Items war wiederum auf einer vierstufigen Skala von 1 für „stimmt gar nicht“ bis 4 für „stimmt völlig“ vorzunehmen. Die empirisch ermittelten Kennwerte für die Skala sind in Tabelle 33 zusammengefasst. Die interne Konsistenz der Skala ist mit Cronbachs $\alpha = .76$ hinreichend hoch.

Tabelle 33 Kennwerte der Skala zur wahrgenommenen Nützlichkeit der Feedbacks

	Nützlichkeit der Feedbacks (5 Items)
N	181 [a]
Mittelwert	2.84
Standardabweichung	0.67
Cronbachs α	.76

Anmerkungen. [a] Kontrollgruppe ausgeschlossen.

Anstrengungsmotivation

Basierend auf den Erfahrungen des ersten Experiments, die auch Zweifel an der Testmotivation der Probanden aufwarfen, wurde in der vorliegenden Untersuchung die Anstrengungsmotivation erfasst. Die Motivation beeinflusst die Bildung der Ziele, die sich die Schüler in Bezug auf die Rezeption der Texte (Guthrie & Wigfield, 2000) und/oder die Verarbeitung der Feedbackmitteilungen setzen (Bangert-Drowns et al., 1991; vgl. auch Abschnitt 3.5.2), und kommt vermittelt über beispielsweise die Dauer und Qualität der Auseinandersetzung mit dem Material, den Einsatz bestimmter Lesestrategien und/oder die Konzentration zum Tragen (Möller & Schiefele, 2004).

Als ein Maß für die Anstrengungs- bzw. Testmotivation wurde auf die intraindividuelle Anstrengungsbereitschaft zurückgegriffen, die mittels des „Anstrengungsthermometers“ aus PISA 2000 (Kunter et al., 2002) gemessen wurde.

> Dabei wird die maximal mögliche Anstrengung als die Anstrengung, die der Schüler in einer für ihn persönlich wichtigen Situation investieren würde, vorgegeben (Wert 10 auf der Skala). Die Schüler sollen angeben, wie sehr sie sich im Vergleich zu ihrer maximalen Anstrengung bei dem gerade bearbeiteten Leistungstest angestrengt haben. (Kunter et al., 2002, S. 208)

Die Antwortskala für das Item war also zehnstufig angelegt, von 1 für „minimale Anstrengung“ bis 10 für „maximale Anstrengung“. Die entsprechende Aufgabenstellung zum Item lautet im Original:

> Stell dir bitte eine Situation (innerhalb oder außerhalb der Schule) ganz konkret vor, die für dich persönlich sehr wichtig ist und in der du dein Bestes geben und dich maximal anstrengen würdest. Wie sehr hast du dich im Vergleich zu der gerade vorgestellten Situation bei diesem Test angestrengt? (Kunter et al., 2002, S. 208)

In der vorliegenden Untersuchung interessierte die Anstrengungsbereitschaft der Probanden in Bezug auf die Bearbeitung der Units in der Treatmentphase. Das Instrument wurde deshalb vor dem Posttest platziert und die Formulierung der Aufgabenstellung des Originalitems („Wie sehr hast du dich …?“) wie folgt angepasst (Änderungen kursiv): „Wie sehr hast du dich im Vergleich zu *dieser* gerade vorgestellten Situation *bei den zuvor bearbeiteten Texten und Aufgaben* angestrengt?“ Das Item wurde computerbasiert dargeboten.

Die in dieser Untersuchung gefundenen empirischen Kennwerte zum Anstrengungsthermometer sind in Tabelle 34 dargestellt.

Tabelle 34 Kennwerte der Skala zur Anstrengungsmotivation

	Anstrengungs-motivation (theoretisches Max: 10)
N	241 [a]
Mittelwert	7.54
Standardabweichung	1.92
Minimum	1
Maximum	10

Anmerkungen. [a] Fallausschlüsse aufgrund vorzeitiger Abbrüche (vgl. Abschnitt 10.8).

10.4 Material (Texte und Items)

Das eingesetzte Material[8] basiert auf den Texten und Items der ersten Untersuchung. Es wurde allerdings auf der Grundlage der Testkennwerte der ersten Untersuchung in Teilen angepasst. Zum einen wurden einige Aufgaben entnommen, und zwar vornehmlich jene, deren Beantwortung ein umfassendes, integriertes Situationsmodell erforderte und sich dabei auf den Text als Ganzes bezogen. Andere Aufgaben wurden geändert, wenn beispielsweise die Häufigkeitsverteilungen der Distraktoren aus dem ersten Experiment auf einen weniger plausiblen Distraktor hinwiesen. Um den Wegfall an Items zu kompensieren, wurden wiederum neue Aufgaben zu den bestehenden Materialien konstruiert. Diese erforderten wieder das Herstellen von Sinnzusammenhängen zwischen Satzteilen oder benachbarten Sätzen oder zwischen mehreren Sätzen und/oder über Abschnitte hinweg. Um diese zusätzlichen Aufgaben entwickeln zu können, war es zum anderen jedoch notwendig, auch die Texte teilweise anzupassen, indem Informationen entweder eingeschoben oder entnommen wurden, so dass aus einer expliziten Aussage eine implizite Information wurde und vice versa. Auf diese Weise wurde Spielraum für neue fünfstufige Multiple-Choice Items des erläuterten Anforderungsbereichs geschaffen.

Des Weiteren wurde eine Unit der Treatmentphase ausgetauscht. Es handelte sich dabei um die Unit „Klimawandel“. Das Thema Klimawandel hatte zum Zeitpunkt des zweiten Experiments eine starke Aktualität in der öffentlichen Diskussion/Berichterstattung erfahren. Es wurde daher vermutet, dass auch die anvisierte Altersgruppe der Sechstklässler im Unterricht oder außerschulisch via verschiedener Medien verstärkt mit dem Thema konfrontiert und damit entsprechendes Vorwissen generiert worden sein könnte, was die Beantwortung der entsprechenden Aufgaben der Unit Klimawandel im Test übervorteilt hätte. Als Ersatz diente eine Unit aus dem Follow-up des ersten Experiments („Der Elefantenrüsselfisch“), die ebenfalls auf einem Sachtext beruhte. Die Feedbacks für die Items dieser Unit wurden für das vorliegende, zweite Experiment entsprechend neu konstruiert.

[8] Die Texte und Aufgaben können bei Bedarf bei der Autorin angefordert werden.

10.5 Prozedur der Feedbackgabe

Die Feedbackgabe verlief nach demselben Prinzip wie in der ersten Untersuchung: Wenn im Experiment eine Aufgabe im ersten Versuch falsch beantwortet wurde, erhielten die Probanden der Experimentalgruppen ein Feedback. Danach war dieselbe Aufgabe ein zweites Mal zu beantworten. Die zuvor gewählte falsche Antwortalternative konnte dabei nicht mehr ausgewählt werden. Nach der zweiten Antwort für ein Item erschien automatisch das nächste Item. Für Aufgaben, die dagegen im ersten Versuch richtig gelöst wurden, gab es kein Feedback; korrekte Erstversuche wurden direkt von der nächsten Aufgabe gefolgt.

Die Probanden der Kontrollgruppe, die entsprechend dem Versuchsplan nie Feedback erhielten, konnten alle Items nur einmal beantworten. Wenn eine Aufgabe beantwortet war, wurde automatisch die nächste präsentiert.

10.6 Das computerbasierte Programm

Das Computerprogramm, mit dem das Experiment und der Posttest administriert wurden, war mit einer Ausnahme dasselbe wie im ersten Experiment. Für die grundlegende Beschreibung seines Aufbaus und der technischen Merkmale wird deshalb auf den Abschnitt 5.6 (S. 105) verwiesen. Neu in dieser Untersuchung war, dass mit jedem Öffnen des Programms eine mehrstellige Ziffer zufällig generiert wurde. Diese Identifikationsnummer war auf der Programmoberfläche sichtbar. Mithilfe der Nummer war jedes Programm und damit jede Ergebnisdatei eineindeutig identifizierbar. Auf das Erfassen von Namen konnte damit verzichtet werden, auch wenn wie in dieser Untersuchung pro Person eine weitere Datenquelle, nämlich das Testheft, der Ergebnisdatei des Computerprogramms zugeordnet werden musste.

In der Experimentalbedingung Inferenzprompt-via-Testleiter blieb das Feedbackfenster der Programmoberfläche leer. Dass eine Aufgabe im ersten Versuch falsch beantwortet worden war, konnte der Testleiter hier zum einen anhand eines auf dem Bildschirm erscheinenden Zeichens (Punkte) erkennen. Zum anderen war in einem zweiten Antwortversuch die zuvor ausgewählte, falsche Antwortalternative in einem deutlich hellen Farbton dargestellt.

10.7 Untersuchungsdurchführung

Die Untersuchung fand während der regulären Unterrichtszeit statt und nahm eine doppelte Unterrichtsstunde, die ohne Unterbrechung durchgeführt wurde, in Anspruch. Die Dauer eventueller Schulpausen während der Untersuchungszeit durfte in Absprache mit dem zuständigen Lehrer bzw. Schulleiter meist der Untersuchungszeit zugeschlagen werden. Die Untersuchung wurde wiederum hauptsächlich durch geschulte, studentische Hilfskräfte durchgeführt. Am Ende der Sitzung erhielten alle Teilnehmer eine kleine Aufmerksamkeit (einen Bleistift) als Dankeschön für ihre Teilnahme.
Zu Beginn der Untersuchungssitzung wurden die Probanden über den Zweck und das Ziel der Untersuchung und den Ablauf der Sitzung aufgeklärt. Danach wurden sie den Versuchsbedingungen randomisiert zugewiesen. Die Experimentalbedingung Inferenzprompt-via-Testleiter wurde in Einzelsitzungen durchgeführt, wozu die Schüler dieser Bedingung zusammen mit den ihnen zugeteilten Testleitern einzelne Räume (z.B. Konferenzraum, leere Klassenräume) aufsuchten. Die Probanden der anderen drei Versuchsbedingungen verblieben im schuleigenen Computerraum und arbeiteten wie alle Teilnehmer des ersten Experiments in der Gruppensitzung.

10.7.1 Gruppensitzung

Zu Beginn wurde das Salzburger Lesescreening, das als Papier-und-Bleistift-basiertes Testheft vorlag, bearbeitet. Im Anschluss daran wendeten sich alle Teilnehmer dem Computerprogramm zu. Die Instruktionen für die Bearbeitung von Fragebogen, Experiment und Posttest erfolgten wie gehabt über das Computerprogramm. So konnte gewährleistet werden, dass die ebenfalls in der Gruppensitzung anwesende Kontrollgruppe, die kein Treatment erhielt, nicht durch die Erläuterungen für die Experimentalgruppen zur Feedbackgabe bei falschen Antworten beeinflusst wurde. Alle Probanden arbeiteten selbstständig am Computer. Der Testleiter war anwesend. Das Übertragen der Identifikationsnummern der Computerprogramme auf die Testhefte der jeweiligen Probanden wurde vom Testleiter vorgenommen.

10.7.2 Einzelsitzung

Die Experimentalbedingung Inferenzprompt-via-Testleiter wurde in Einzelsitzungen durchgeführt, da das Sprechen der Testleiter andere Probanden gestört hätte. Für die Einzelsitzungen waren in den teilnehmenden Schulen parallel zur Gruppensitzung weitere, freistehende Räume (z.B. Elternsprechzimmer, Lehrerbibliothek, Beratungszimmer) zur Verfügung gestellt worden. Für die Durchführung des Computerprogramms wurde dabei auf Laptops zurückgegriffen.

Der Ablauf der Sitzung war analog zur Gruppensitzung gestaltet. Abweichungen ergaben sich an den folgenden Stellen: Für die Dauer des computerbasierten Tests saß der Testleiter neben dem Probanden. Zu Beginn des Computerprogramms erläuterte der Testleiter die Rolle, die er während der Bearbeitung des Tests einnehmen würde. Die Instruktionen zum Ablauf und zur Bedienung des Programms wurden jedoch wie in den anderen Versuchsgruppen computerbasiert gegeben. Wenn im Verlauf des Experiments eine Feedbackgabe angezeigt war, trug der Testleiter die Rückmeldung mündlich vor und legte danach eine entsprechende Karte, auf der dieselbe Information stand, für den Probanden gut sichtbar neben den Laptop. Sobald der Proband die zweite Antwort für das Items abgegeben hatte, nahm der Testleiter die Karte wieder weg. Die Feedbacks waren identisch mit denen der Versuchsbedingung Inferenzprompt via Computer. Der Testleiter gab keine zusätzlichen Informationen, auch keine bewussten, nonverbal kommunizierten Rückmeldungen. Das Feedback wurde in natürlicher Sprechart vorgetragen. Die zusätzlich eingesetzten Feedbackkarten dienten der Unterstützung des Informationsverarbeitungsprozesses, da durch sie die Überarbeitung der Aufgabenstellung weniger von den Erinnerungsleistungen an die Rückmeldung beeinflusst wurde.

10.8 Datenanalyse

10.8.1 Fehlende Werte und Fallausschluss

Der als Papier-und-Bleistift-Verfahren umgesetzte Lesegeschwindigkeitstest wurde von allen N = 251 Untersuchungsteilnehmern bearbeitet. Somit liegen hierfür zunächst keine fehlenden Gesamttestwerte vor. In zwei Fällen blieb der Test aufgrund instruktionswidriger Bearbeitung für die Analysen jedoch unberücksichtigt. Daraus ergibt sich hier eine Auswertungsstichprobe von N = 249.

Für die computeradministrierten Lesekompetenztests der Treatmentphase bzw. des Posttests und Fragebogen treten fehlende Werte dann auf, wenn das Programm nicht weiter bearbeitet wurde. In insgesamt 15 Fällen liegt kein vollständiger Datensatz vor. Davon weist in acht Fällen bereits das Experiment fehlende Werte auf und in vier Fällen ist der Posttest angefangen, aber nicht beendet. Die restlichen drei Fälle haben die Untersuchung beim Fragebogenteil zwischen Experiment und Posttest beendet.
Es zeigt sich, dass die unvollständigen Datensätze vermehrt aus der Versuchsbedingung Inferenzprompt-via-Testleiter stammen (10 der 15 Fälle). Dieser Umstand ist primär darin begründet, dass das Aufsuchen der Räume für die in dieser Versuchsbedingung notwendigen Einzelsitzungen bei organisatorischen Komplikationen in den Schulen mitunter zu deutlichen Verzögerungen des Testbeginns führte. Die fehlenden Werte in der Testleiterbedingung und auch den anderen Versuchsgruppen wurden bei der Summenscorebildung nicht als Fehler gewertet und die unvollständigen Datensätze aus Experiment oder Posttest wurden aus den jeweiligen Analysen ausgeschlossen.

Des Weiteren wurden wie im ersten Experiment Fälle ausgeschlossen, in denen zu viele Lesekompetenzitems in zu kurzer Zeit abgeschlossen („durchgeklickt“) wurden, weil hier eine ernsthafte Testbearbeitung in Frage gestellt werden kann. Als Cut-off-Kriterium, unter dem ein Item als „zu schnell beantwortet“ eingestuft wurde, galt wiederum der Wert von sechs Sekunden. Jene Fälle, die mehr als fünf der insgesamt 35 Items des Lesekompetenztests in der Treatmentphase (= 14.3 % des Testteils) bzw. mehr als zwei der insgesamt 14 Posttestitems (= 14.3 % des Testteils) in weniger als sechs Sekunden abgeschlossen hatten, wurden aus den Analysen des Experiments bzw. Posttests ausgeschlossen. Die gewählten Cut-off-Werte bezüglich der Schnelligkeit der Beantwortung von Items und der Menge zugelassener „durchgeklickter“ Items sind vom ersten Experiment übernommen; für zusätzliche Erläuterungen zum Vorgehen und die Begründung der Kriterien wird dementsprechend auf Abschnitt 5.8.1 verwiesen.
Tabelle 35 enthält die Häufigkeitsverteilung bezüglich des genutzten Kriteriums von weniger als sechs Sekunden. Daneben sind ebenfalls die Häufigkeitsverteilungen von drei alternativen, höheren Cut-off-Werten (vgl. Abschnitt 5.8.1) abgebildet. Diese alternativen Kriterien (und zwar < 7 / < 8 / < 9 Sekunden) wurden nicht für Fallausschlüsse

herangezogen[9]. Die Gegenüberstellung verdeutlicht aber, dass diese Kriterien in der vorliegenden Untersuchung meist zu wenigen Fallausschlüssen mehr geführt hätten als durch die Wahl des genutzten Cut-off-Wertes (6 Sek.) geschehen.

Die Häufigkeitsverteilungen in Tabelle 35 zeigen zudem, dass wiederum die (große) Mehrheit der Teilnehmer für alle Items mehr Zeit bis zu deren Beantwortung aufbrachte als durch die jeweiligen kritischen Werte festgelegt. Das heißt, die meisten Probanden hatten kein einziges Item „durchgeklickt“, gleichgültig welcher Cut-off-Wert zugrunde gelegt wurde. Außerdem kamen einzelne durchgeklickte Items eher vor als viele durchgeklickte Items pro Person.

Aufgrund des genutzten Kriteriums (weniger als sechs Sekunden für eine Antwort) waren für das Experiment 13 Fälle und für den Posttest 27 Fälle auszuschließen. Unter Berücksichtigung der Bedingung, dass nur jeweils vollständige Datensätze in die Analysen einbezogen werden, ergeben sich damit als Auswertungsstichproben N = 230 für den Lesekompetenztest im Experiment und N = 209 für den Posttest.

10.8.2 Scoring

Die Antworten im Lesekompetenztest des Experiments und des Posttests wurden stets als richtig/falsch mit der Codierung 1/0 gewertet. Die ersten und zweiten Versuche der Experimentalgruppen im Experiment wurden dabei gleichwertig behandelt – eine richtige Antwort im zweiten Versuch wurde ebenfalls mit 1 gewertet. Die Einzelantworten wurden zu Summenscores aggregiert. Somit ergab sich hier für alle Versuchsgruppen ein Summenwert für die ersten Versuche und für die Experimentalgruppen zusätzlich ein Summenwert für die zweiten Versuche. Der Summenwert der zweiten Versuche steht allerdings immer in Abhängigkeit vom Summenwert der ersten Versuche, denn je nachdem wie viele Fehler in den Erstantworten gemacht wurden, ergab sich so die Anzahl zweiter Versuche. Weiterhin konnten alle Versuchsgruppen in den Summenscores für den Posttest verglichen werden.

Zur Auswertung der Feedbackeffekte kamen hauptsächlich varianzanalytische Verfahren zum Einsatz, die mit der Software SPSS ausgeführt wurden.

---

[9] Die zentralen Analysen dieses Experiments wurden zusätzlich unter der Nutzung der drei alternativen Cut-off-Werte durchgeführt. Die Ergebnisse dieser Analysen (siehe Anhang C) widersprechen nicht den im Hauptteil der Arbeit berichteten Ergebnissen.

Tabelle 35 Häufigkeit „durchgeklickter“ Items

	Menge der Items [a]	Cut-off-Werte, *unter denen* ein Item als „zu schnell beantwortet“ gilt											
		< 6 Sekunden (= gewähltes Kriterium)			< 7 Sekunden			< 8 Sekunden			< 9 Sekunden		
		N	%		N	%		N	%		N	%	
Test der Treatmentphase (N = 35 Items)	0	210	83.7	Σ = 238	201	80.1	Σ = 234	190	75.7	Σ = 231	157	62.5	Σ = 231
	1	13	5.2		17	6.8		23	9.2		39	15.5	
	2	3	1.2		8	3.2		8	3.2		20	8.0	
	3	6	2.4		2	0.8		6	2.4		5	2.0	
	4	1	0.4		3	1.2		4	1.6		6	2.4	
	5	5	2.0		3	1.2		0	0		3	1.2	
	6	3	1.2	Σ = 13	7	2.8	Σ = 17	6	2.4	Σ = 20	5	2.0	Σ = 21
	7	0	0		0	0		2	0.8		3	1.2	
	8	1	0.4		0	0		2	0.8		1	0.4	
	9	2	0.8		2	0.8		1	0.4		3	1.2	
	10	1	0.4		1	0.4		2	0.8		2	0.8	
	>10	5	2.0		6	2.4		6	2.4		6	2.4	
	>20	1	0.4		1	0.4		1	0.4		1	0.4	
	>30	0	0		0	0		0	0		0	0	
	Σ	251	100		251	100		251	100		251	100	
Posttest (N = 14 Items)	0	204	81.3	Σ= 224	186	74.1	Σ= 222	168	66.9	Σ= 217	142	56.6	Σ= 205
	1	16	6.4		29	11.6		32	12.7		46	18.3	
	2	4	1.6		7	2.8		17	6.8		17	6.8	
	3	11	4.4	Σ = 27	6	2.4	Σ = 29	8	3.2	Σ = 34	14	5.6	Σ = 43
	4	0	0		6	2.4		7	2.8		8	3.2	
	5	4	1.6		3	1.2		3	1.2		5	2.0	
	6	3	1.2		4	1.6		4	1.6		7	2.8	
	7	0	0		1	0.4		3	1.2		0	0	
	8	1	0.4		1	0.4		0	0		0	0	
	9	0	0		0	0		0	0		0	0	
	10	2	0.8		1	0.4		2	0.8		0	0	
	>10	6	2.4		7	2.8		7	2.8		9	3.6	
	Σ	251	100		251	100		251	100		251	100	

Anmerkungen. [a] In der Spalte sind die Anzahl der Items, auf die das jeweilige Cut-off-Kriterium zutrifft, abgetragen (d.h. wie oft wurden Items unterhalb der zeitlichen Grenzwerte beantwortet).

10.8.3 Auswertungsplan

Der Auswertungsplan zum vorliegenden Experiment spiegelt das Vorgehen bei der Analyse des ersten Experiments wider und berücksichtigt dabei im Wesentlichen folgende Schritte: Zunächst wurde die „Qualität“ der in der Untersuchung eingesetzten Lesekompetenztests untersucht. Diese erneute Analyse war notwendig, weil das Material im Vergleich zur ersten Untersuchung in Teilen angepasst wurde (vgl. Abschnitt 10.4). Angewendet wurden Item- bzw. Skalenanalysen im Rahmen der Klassischen Testtheorie sowie Raschanalysen, letztere um Hinweise zur Dimensionalität des Itemmaterials zu erhalten. Das Ziel der Analysen besteht wiederum darin, die administrierten Items in ihren Kennwerten zu beschreiben und hinsichtlich der Bewertungskriterien gänzlich ungeeignete Items zu selektieren. Für die Item- bzw. Skalenanalysen werden nur die Daten der interventionsfreien Kontrollgruppe (N = 53) herangezogen (vgl. Abschnitt 10.4).

Im Anschluss an die Analysen der Lesekompetenzitems wurden die Versuchsgruppen hinsichtlich der zur Verfügung stehenden Person- bzw. Hintergrundvariablen auf Gleichverteilung getestet. Diese Überprüfung ergänzt die vorgenommene Randomisierung, die in den jeweilig untersuchten Gruppen/Klassenverbänden eines Untersuchungstermins durchgeführt ist.

Danach wurden die Fragestellungen dieses Experiments bezüglich der Wirksamkeit der Feedbackinterventionen untersucht. Dabei wurden nacheinander zunächst die Feedbackeffekte auf die Leistung in der Treatmentphase und dem Posttest analysiert. Im Anschluss daran wurden die Bearbeitungszeiten und mögliche Effekte der Feedbackinterventionen darauf untersucht. Zur Auswertung der Feedbackeffekte kamen hauptsächlich varianzanalytische Verfahren zum Einsatz, die mit der Software SPSS ausgeführt wurden.

11 Ergebnisse

11.1 Beschreibung der Lesekompetenzitems

Basierend auf den Daten der Kontrollgruppe (N = 53) wurden zunächst alle Lesekompetenzitems der Treatmentphase und des Posttests (insgesamt N = 49 Items) mittels Item- und Reliabilitätsanalysen der Klassischen Testtheorie analysiert. Die Ergebnisse zeigen, dass fast alle Items (N = 48 Items) im mittleren Schwierigkeitsbereich von p = .20 bis p = .80 liegen (Bühner, 2006). Lediglich ein Item weist eine extreme Schwierigkeit von p = .87 auf (Item Spr2, vgl. Tabelle 36). Die mittlere Itemschwierigkeit beträgt p = .51. Die interne Konsistenz ist mit Cronbachs α = .78 (Präzision P_{α} < .01) akzeptabel.

Hinsichtlich der Itemtrennschärfen gibt es wiederum Auffälligkeiten: vier der 49 Items weisen negative Trennschärfen auf und für 12 weitere Items ergeben sich mit r < .19 sehr niedrige Zusammenhänge (Ebel, 1979) zu den restlichen Lesekompetenzitems. Die inhaltliche Analyse dieser Items zeigte in Abgrenzung zu den anderen Items keine spezifischen Abweichungen in den Fragestellungen bzw. der Art der gestellten Anforderung. Sie fragen tendenziell aber eher komplexere Sachverhalte ab, für die die Schüler der Stichprobe möglicherweise nicht zuverlässig eine passende Vorstellung haben aufbauen können.

Die Items mit niedrigen, aber positiven Trennschärfen werden letztlich im Itempool belassen. Die vier Items mit negativen Trennschärfen (Items Ant2, Ant7, Spr4 und Nat2), die aus dem Testteil der Treatmentphase stammen, werden dagegen ausgeschlossen. Auch wenn sie inhaltlich zu den formulierten Anforderungen passen, stehen sie dennoch in einem negativen Zusammenhang zu den restlichen Items und würden daher die zu bildenden Leistungssummenscores eher negativ beeinflussen.

Die verbliebenen 45 Items aus der Treatmentphase und dem Posttest (nunmehr 31 bzw. 14 Items) wurden einer erneuten Reliabilitätsanalyse unterzogen. Die mittlere Itemschwierigkeit liegt mit weiterhin p = .51 im mittleren Bereich (vgl. Tabelle 36 für Schwierigkeiten der einzelnen Items). Es verbleiben 12 Items mit sehr niedrigen Trennschärfen (r < .19; vgl. Tabelle 36). Dennoch stieg Cronbachs alpha durch den

Itemausschluss erwartungsgemäß an und beträgt nun α = .81. Mit einer Präzision von $P_{\alpha} < .01$ spricht der Wert für eine gute interne Konsistenz der verbliebenen Lesekompetenzitems.

Tabelle 36 Itemkennwerte

Nr	Label	δ	MNSQ	T	p	r
			Experiment			
1	Ant1	0.88	0.98	-0.1	.30	0.26
2	Ant3	-0.09	1.12	1.3	.52	0.10
3	Ant4	0.28	1.15	1.6	.43	0.04
4	Ant5	-1.24	1.01	0.1	.78	0.22
5	Ant6	0.28	1.02	0.2	.43	0.24
6	Ele1	0.77	1.05	0.4	.35	0.14
7	Ele2	0.38	0.94	-0.6	.43	0.34
8	Ele3	0.57	0.84	-1.5	.39	0.50
9	Ele4	0.00	1.12	1.2	.50	0.06
10	Ele5	-0.38	1.11	1.1	.59	0.10
11	Ele6	0.77	1.03	0.3	.35	0.18
12	Spr1	-1.12	0.91	-0.5	.76	0.43
13	Spr2	-2.02	1.00	0.1	.87	0.22
14	Spr3	0.67	1.15	1.3	.35	0.03
15	Spr5	-1.38	0.95	-0.2	.78	0.31
16	Spr6	-0.89	1.04	0.3	.70	0.21
17	Spr7	-0.89	0.92	-0.5	.70	0.39
18	Som1	-0.38	1.05	0.5	.61	0.16
19	Som2	0.47	0.96	-0.4	.39	0.34
20	Som3	0.67	1.03	0.3	.37	0.21
21	Som4	0.09	0.99	-0.2	.48	0.32
22	Som5	-0.19	0.96	-0.4	.57	0.31
23	Som6	0.37	1.02	0.2	.43	0.22
24	Som7	-0.48	1.09	0.8	.61	0.09
			Noch Experiment			
25	Nat1	1.47	0.83	-0.9	.22	0.56
26	Nat3	-0.58	0.95	-0.4	.63	0.32
27	Nat4	-0.48	1.00	0.0	.63	0.28
28	Nat5	-0.01	1.11	1.2	.50	0.11
29	Nat6	0.47	0.99	-0.1	.41	0.27
30	Nat7	-0.01	1.05	0.5	.52	0.21
31	Nat8	1.47	0.98	-0.1	.22	0.26
			Posttest			
1	Man1	-0.01	0.91	-1.0	.52	0.44
2	Man2	-0.48	0.91	-0.8	.63	0.41
3	Man3	-0.86	0.94	-0.4	.70	0.38
4	Man4	-0.44	0.89	-1.1	.61	0.48
5	Man5	-0.05	0.99	-0.1	.52	0.30
6	Man6	0.95	0.94	-0.4	.30	0.37
7	Man7	-0.54	0.85	-1.3	.63	0.52
8	Kak1	0.05	1.05	0.6	.50	0.20
9	Kak2	-0.05	1.12	1.3	.52	0.10
10	Kak3	1.07	0.97	-0.2	.28	0.28
11	Kak4	0.34	0.96	-0.3	.43	0.34
12	Kak5	0.05	1.12	1.3	.50	0.13
13	Kak6	-0.44	1.05	0.5	.61	0.22
14	Kak7	0.96	0.98	-0.1	.30	0.29

Anmerkungen. Label = gibt die Unit (vgl. Tabelle 13) und die Nummer des Items darin an; δ = Itemschwierigkeit (Raschmodell); MNSQ = Weighted Mean Square (Raschmodell); T = Wert aus T-Verteilung (Raschmodell); p = Itemschwierigkeit (klassische Testtheorie); r = Itemtrennschärfe (klassische Testtheorie).

In einem weiteren Schritt wurden die verbliebenen 45 Lesekompetenzitems aus Treatment und Posttest im Rahmen der Raschskalierung mit Conquest (Wu et al., 2007) überprüft. Diese zeigt, dass keines der Items nach den Richtwerten von Adams (2002) auffällig abweichende MNSQ-Kennwerte von $1.20 <$ MNSQ < 0.80 und gleichzeitig auffällige T-Werte von $2.0 < T < -2.0$ aufweist. Die entsprechenden Itemkenntwerte aus dem Raschmodell sind in Tabelle 36 aufgeführt und um die Kennwerte der Itemschwierigkeit und Itemtrennschärfe aus der Reliabilitätsanalyse nach der Klassischen

Testtheorie ergänzt. Abbildung 14 zeigt die latente Verteilung der Personen- und der Itemparameter auf einer gemeinsamen Logit-Skala.

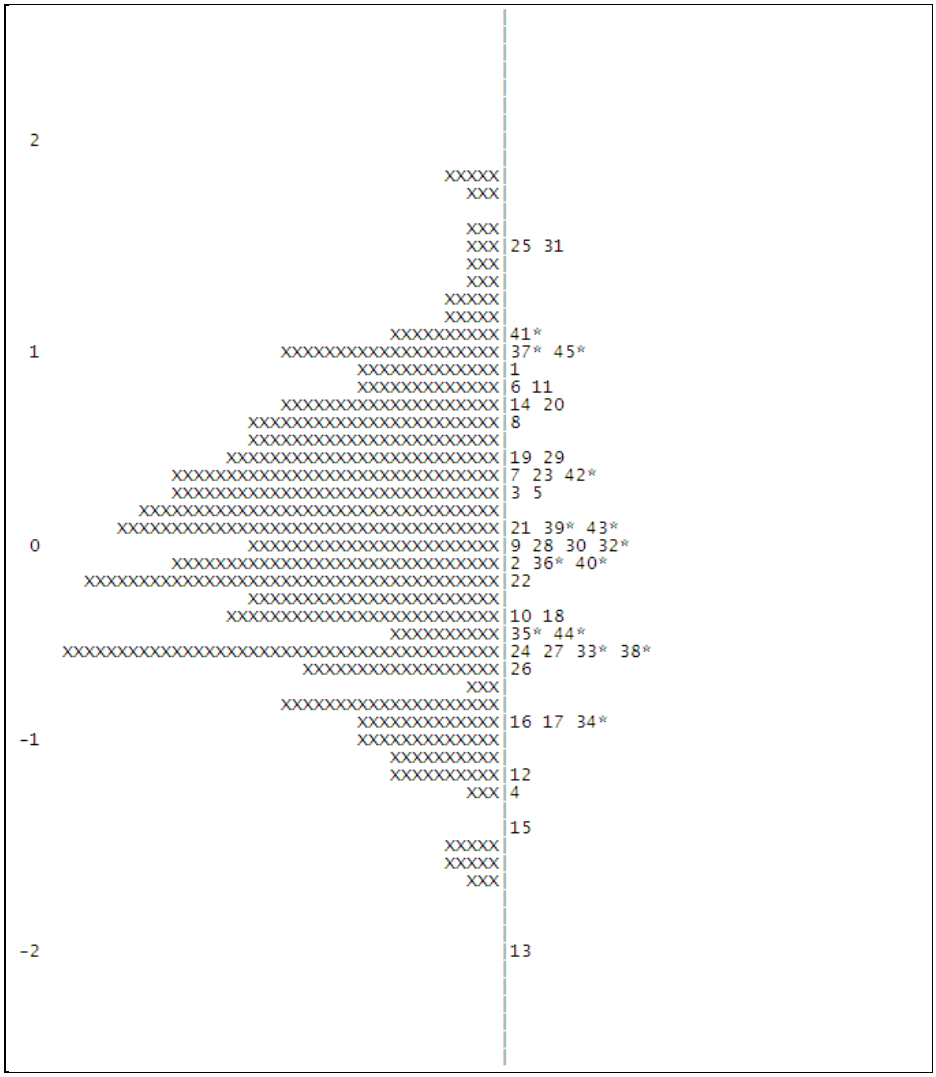

Abbildung 14 Latente Verteilung der Personenparameter (Kreuze links) und Itemparameter (Zahlen rechts) auf einer gemeinsamen Logit-Skala.

Anmerkungen. Jedes 'x' repräsentiert 0.1 Fälle; Die mit * gekennzeichneten Items gehören zum Posttest; Die Itemlabel zu den hier abgebildeten Itemnummern sind der Tabelle 36 zu entnehmen.

Die Ergebnisse der Analysen nach der Klassischen Testtheorie und die Raschanalyse zusammengenommen widersprechen nicht der Annahme, dass die Lesekompetenzitems ein eindimensionales Konstrukt widerspiegeln. Die Leistungen im Experiment und im Posttest wurden in jeweils einem Summenwert zusammengefasst und entsprechend

ausgewertet. Darüber hinaus zeigen die Analysen, dass die Testitems eine im Allgemeinen angemessene Schwierigkeit aufweisen.

11.2 Überprüfung von a-priori Gruppenunterschieden

Die randomisierte Zuweisung der Untersuchungsteilnehmer wurde zusätzlich durch die Überprüfung möglicher a-priori Gruppenunterschiede in leistungsrelevanten Hintergrundvariablen abgesichert. Als Hintergrundvariablen standen die Lesegeschwindigkeit, die Testangst sowie die Personmerkmale Geschlecht und Herkunftssprache zur Verfügung. Die versuchsgruppenbezogenen Mittelwerte und Standardabweichungen im Lesegeschwindigkeitstest und dem Fragebogen zur Testangst sind in Tabelle 37 dargestellt. Die varianzanalytischen Auswertungen beider Instrumente zeigen, dass sich die Versuchsgruppen a-priori weder in den Ausprägungen der Lesegeschwindigkeit (F (3, 224) = 0.29; $p > .05$; $\eta^2 = .004$) noch der Testangst (F (3, 226) = 0.16; $p > .05$; $\eta^2 = .002$) unterschieden.

Tabelle 37 Deskriptive Statistiken zur Lesegeschwindigkeit und Testangst

		N	*M*	*SD*
Lesegeschwindigkeit (N = 228)	Kein Feedback	53	40.23	6.96
	Knowledge of Result	57	40.58	7.14
	Inferenzprompt	59	40.53	7.03
	Inferenzprompt-via-Testleiter	59	39.51	7.00
Testangst (N = 230)	Kein Feedback	53	1.97	0.56
	Knowledge of Result	58	1.98	0.50
	Inferenzprompt	60	2.03	0.62
	Inferenzprompt-via-Testleiter	59	2.03	0.49

Tabelle 38 Häufigkeiten für Geschlecht und Herkunftssprache (N = 230)

	Knowledge of Result	Inferenz-prompt	Inferenz-prompt-via-Testleiter	Kein Feedback	Σ
Geschlecht					
Mädchen	31	33	23	27	114
Jungen	27	27	36	26	116
Σ	58	60	59	53	230
Sprache [a]					
Deutsch	54	59	56	51	220
andere	4	1	3	2	10
Σ	58	60	59	53	230

Anmerkungen. [a] Die Angabe zur Sprache bezieht sich auf die Frage, welche Sprache das Kind hauptsächlich zu Hause spricht.

Die Analyse des Mädchen- und Jungenanteils mittels des Chiquadrat-Tests belegt, dass beide Geschlechter über die Versuchsgruppen gleichverteilt waren ($\chi^2 = 3.74$; $p > .05$). Zudem ergab der Chiquadrat-Test bezüglich der Sprache, die die Schüler ihren Angaben nach hauptsächlich zu Hause sprechen, dass Deutsch und nicht deutsche Sprachen zwischen den Versuchsgruppen gleichverteilt waren ($\chi^2 = 2.06$; $p > .05$). Die Häufigkeiten für die Variablen Geschlecht und Sprache sind in Tabelle 38 zusammengefasst.

Für die vier verfügbaren Hintergrundvariablen konnten damit zufällig aufgetretene a-priori Gruppenunterschiede, die sich in den nachfolgenden Analysen zur Feedbackwirksamkeit möglicherweise als Quelle für Leistungsunterschiede erwiesen hätten, ausgeschlossen werden.

11.3 Haupteffekte von Feedback auf die Leistung

11.3.1 Die Leistung in der Treatmentphase

Erstantworten

Zur Beantwortung der zentralen Fragestellung nach den Feedbackeffekten auf das Textverständnis/die Lesekompetenz wurden zunächst die Leistungen in den Erstantworten der Treatmentphase betrachtet. Hier zeigt die deskriptive Statistik (vgl. Tabelle 39), dass die Bedingung Inferenzprompt-via-Testleiter im Mittel die höchste

Leistung hervorbrachte, die Bedingung der computerbasierten Darbietung der Inferenzprompts dagegen den niedrigsten Mittelwert. Die mittleren Leistungen der beiden Kontrollbedingungen liegen dazwischen. Die ANOVA belegt einen signifikanten Haupteffekt des Feedbacks auf die Leistung in den Erstantworten (F (3, 226) = 4.23; $p < .01$; $\eta^2 = .05$).

Der signifikante Effekt geht, wie die Scheffé-Tests zeigen, auf den Unterschied zwischen den beiden Bedingungen mit Inferenzprompts zurück, und zwar zu Gunsten der Bedingung Inferenzprompt-via-Testleiter: die mittlere Differenz beträgt hier $\Delta = 3.21$ ($p < .01$). Dieser Leistungsvorteil ist nach Cohen (1992) mit $d = 0.64$ (Cohens d) als mittlerer Effekt einzuschätzen.

Im Kontrast zu den beiden Kontrollbedingungen bringt die Gruppe Inferenzprompts-via-Testleiter keinen statistisch signifikanten Leistungsvorteil hervor. Der Unterschied zur Kontrollbedingung ohne Feedback beträgt im Mittel $\Delta = 1.57$ ($p > .05$; Cohens $d = 0.33$) und zur Bedingung Knowledge of Result im Mittel $\Delta = 2.04$ ($p > .05$; Cohens $d = 0.43$).

Zudem weist die deskriptive Statistik, die auf der Anzahl richtiger Antworten basiert, implizit darauf hin, dass bei den Erstantworten der insgesamt 31 Items im Durchschnitt ungefähr 14 bis 17 Fehler gemacht wurden. Das entspricht einer durchschnittlichen Fehlerquote von 44 % bis 54 %. Die Feedbackbedingungen erhielten somit im Durchschnitt also etwa 14 bis 17 Rückmeldungen im Verlauf der Treatmentphase.

Tabelle 39 Deskriptive Statistik der Erstantworten in der Treatmentphase (N = 230)

	Leistung in den Erstantworten (31 Items)			
	N	*M*	*SD*	*M* %
Kein Feedback	53	15.85	5.02	51.13
Knowledge of Result	58	15.38	4.93	49.61
Inferenzprompt	60	14.22	5.45	45.87
Inferenzprompt-via-Testleiter	59	17.42	4.47	56.19

Um den signifikanten Haupteffekt des Feedbacks auf die Leistung in den Erstantworten näher zu beleuchten, wurde die Leistung in den Erstantworten zusätzlich pro Unit ausgewertet. Sollte sich die Mehrleistung ausschließlich am Anfang der Treatmentphase zeigen, wären die Ergebnisse anders zu interpretieren, als wenn sich die Mehrleistung etwa kontinuierlich über den Verlauf der Treatmentphase nachweisen ließe. Als

Auswertungseinheiten dienten hier die fünf Units der Lesekompetenztests. Um eine bessere Vergleichbarkeit zwischen den Units herzustellen, wurde die Anzahl richtiger Antworten pro Unit an der Anzahl der Items der jeweiligen Unit relativiert.

Die Auswertung mittels MANOVA ergab einen statistisch signifikanten Effekt der Versuchsgruppen (F (3, 226) = 1.85; $p < .05$; $\eta^2 = .04$). In den ersten beiden Units sind keine statistisch signifikanten Unterschiede zwischen den Gruppen zu verzeichnen (vgl. Tabelle 40 für deskriptive Statistiken und F-Werte). In der dritten und der fünften Unit zeigt sich jeweils ein signifikanter Unterschied zwischen den Versuchsgruppen mit Effektstärken von $\eta^2 = .06$ bzw. $\eta^2 = .05$. Die Leistungsunterschiede in der vierten Unit sind auf den Niveau von $\alpha = 0.10$ als tendenziell signifikant und mit einem kleinen Effekt von $\eta^2 = .03$ zu beschreiben. Die signifikanten Effekte gehen jeweils auf den Kontrast zwischen den Bedingungen Inferenzprompt und Inferenzprompt-via-Testleiter zurück. Abbildung 15 illustriert die Mittelwertsunterschiede zwischen den Gruppen.

Tabelle 40 MANOVA für Erstantworten pro Unit (N = 230)

		Leistung in den Erstantworten (relative Lösungshäufigkeit pro Unit [a])				
		Unit 1	Unit 2	Unit 3	Unit 4	Unit 5
	N	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*
Kein Feedback	53	0.52 (0.24)	0.44 (0.24)	0.69 (0.25)	0.49 (0.24)	0.43 (0.24)
Knowledge of Result	58	0.46 (0.25)	0.46 (0.24)	0.64 (0.27)	0.50 (0.21)	0.43 (0.24)
Inferenzprompt	60	0.48 (0.29)	0.39 (0.21)	0.57 (0.28)*	0.47 (0.25)*	0.39 (0.24)*
Inferenzprompt-via-Testleiter	59	0.48 (0.23)	0.49 (0.25)	0.73 (0.20)*	0.58 (0.23)*	0.53 (0.21)*
		$F = 0.66$ $p > .05$ $\eta^2 = .01$	$F = 1.69$ $p > .05$ $\eta^2 = .02$	$F = 4.77$ $p < .01$ $\eta^2 = .06$	$F = 2.25$ $p < .10$ $\eta^2 = .03$	$F = 3.72$ $p < .05$ $\eta^2 = .05$

Anmerkungen. Die Werte basieren auf den richtigen Antworten im Erstversuch, die zum Zweck der Vergleichbarkeit jeweils an der Anzahl der Items der Unit relativiert wurden.

* kennzeichnet die Gruppen, deren mittlere Differenz statistisch signifikant ist.

[a] Die abgebildeten Werte beziehen sich auf die Summenscores pro Unit, die an der Anzahl der Items der jeweiligen Unit relativiert sind.

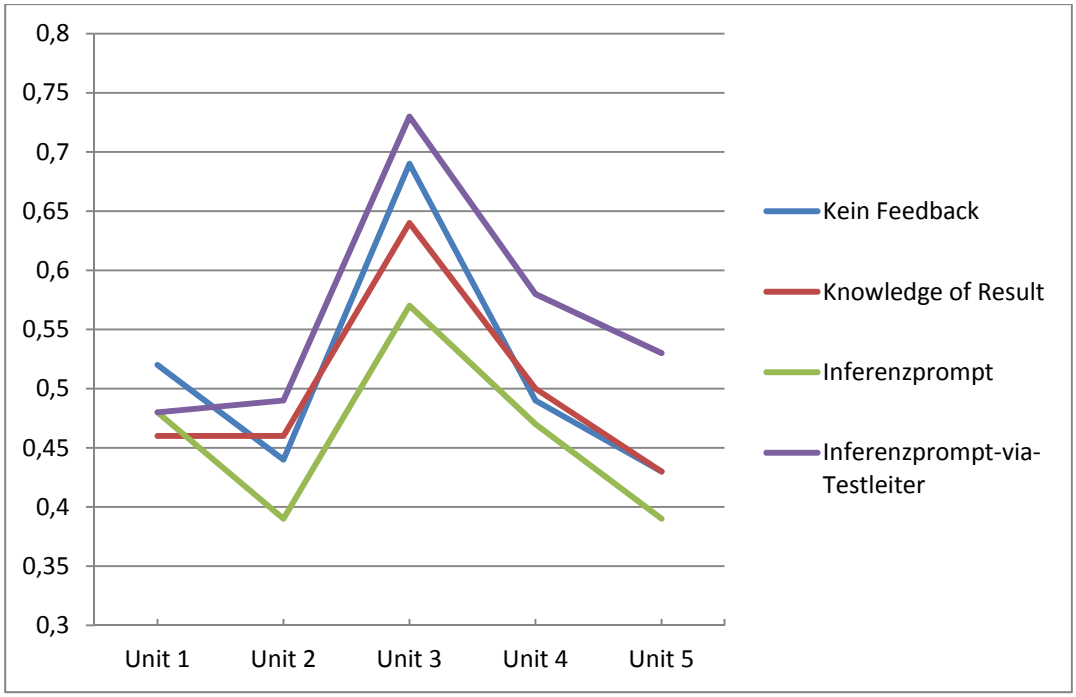

Abbildung 15 Relative Lösungshäufigkeiten in den Erstversuchen, pro Unit.

Zweitantworten

Für die Beurteilung der Feedbackwirksamkeit wurden in den Feedbackbedingungen neben den Erstantworten auch die Leistungen in den zweiten Versuchen herangezogen. Die Mittelwerte und Standardabweichungen der Zweitantworten sind in Tabelle 41 wiedergegeben. Dabei sind für den Gruppenvergleich die relativen Lösungshäufigkeiten entscheidend, da die Anzahl der Zweitversuche von der Anzahl der Fehler in den Erstversuchen abhing.

Auf der deskriptiven Ebene zeigt sich, dass die Bedingung Inferenzprompt-via-Testleiter mit einem Mittelwert von $M = 0.49$ die durchschnittlich höchste Leistung erbrachte und damit etwa jede zweite Aufgabe, die im ersten Versuch falsch beantwortet wurde, nach der Feedbackmitteilung richtig gelöst wurde. In den beiden anderen Feedbackbedingungen, in denen die Rückmeldungen rein computerbasiert dargeboten wurden, gelang die Korrektur falscher Erstantworten im Durchschnitt mit $M = 0.40$ (Knowledge of Result) bzw. $M = 0.41$ (Inferenzprompt).

Tabelle 41 Deskriptive Statistik der Zweitantworten in der Treatmentphase (N = 177)

		Leistung in den Zweitantworten				
		Absolute Häufigkeiten		Relative Häufigkeit [a]		
	N	*M*	*SD*	*M*	*SD*	*M* %
Knowledge of Result	58	6.04	2.46	0.40	0.15	39.78
Inferenzprompt	60	6.30	2.27	0.41	0.16	41.41
Inferenzprompt-via-Testleiter	59	6.41	2.23	0.49	0.17	49.17

Anmerkungen. [a] Die Summe richtiger Antworten im 2. Versuch ist relativiert an der Anzahl der benötigten 2. Versuche.

Die varianzanalytische Auswertung der relativen Lösungshäufigkeiten erbrachte einen signifikanten Effekt des Feedbacks auf die Leistung in den zweiten Versuchen (F (2, 174) = 6.62; $p < .01$; $\eta^2 = .07$). Die Post-hoc Tests (Scheffé) belegen, dass dieser Haupteffekt auf den Leistungsvorsprung der Bedingung Inferenzprompt-via-Testleiter sowohl gegenüber der Versuchsbedingung Inferenzprompt (mittlere $\Delta = 0.9$; $p < .01$) als auch gegenüber Knowledge of Result (mittlere $\Delta = 0.9$; $p < .01$) zurückgeht.

Die Mehrleistung der Testleiterbedingung macht dabei durchschnittlich je 9 % aus. Dieser Unterschied drückt sich in Effektstärken von Cohens $d = 0.56$ (Knowledge of Result) und Cohens $d = 0.49$ (Inferenzprompt) aus, die jeweils als mittlerer Effekt zu bewerten sind. Die beiden rein computerbasierten Feedbackbedingungen Inferenzprompt und Knowledge of Result unterscheiden sich nicht statistisch bedeutsam voneinander (mittlere $\Delta = 0.00$; $p > .05$; $d = |0.07|$).

Gemeinsame Betrachtung der Erst- und Zweitantworten

Der Effekt des Leistungsvorteils der Bedingung Inferenzprompt-via-Testleiter wird durch das Verknüpfen der Leistungen in den Zweit- und Erstversuchen zusätzlich verdeutlicht (vgl. Abbildung 16). Wird jede richtige Antwort im Lesekompetenztests des Treatments, egal ob diese im ersten oder im zweiten Versuch erbracht wurde, gezählt und gleich gewichtet, zeigt sich das folgende Bild: die Testleiterbedingung löst im Durchschnitt $M = 23.83$ Aufgaben ($SD = 3.62$), die Bedingung Knowledge of Result bewältigt $M = 21.41$ Aufgaben ($SD = 4.40$) und die Gruppe Inferenzprompt erreicht $M = 20.52$ Punkte ($SD = 4.97$). Diese durchschnittlichen Summenwerte entsprechen 76.87 % (Inferenzprompts-via-Testleiter), 69.07 % (Knowledge of Result) bzw. 66.19 %

(Inferenzprompt) richtig gelöster Aufgaben im Lesekompetenztest der Treatmentphase. Der Vergleich zur Testleistung der Kontrollbedingung ohne Feedback (M = 15.85; SD = 5.02; 51.13 %) ist an dieser Stelle nicht belastbar, da die Kontrollbedingung schließlich nie zweite Antwortversuche eingeräumt bekam.

Die Unterschiede der Feedbackbedingungen im Gesamttestwert sind, nach den Ergebnissen der separaten Betrachtung von Erst- und Zweitantworten, erwartungsgemäß statistisch signifikant (F (2, 174) = 9.13; $p < .001$; $\eta^2 = .10$). Dabei ist es wiederum die Testleiterbedingung, die sich den Post-hoc-Tests (Scheffé) zufolge statistisch signifikant von den beiden anderen Bedingungen abhebt (zur Gruppe Inferenzprompt mit mittlerer $\Delta = 3.31$; $p < .001$; $d = 0.69$ und zur Gruppe Knowledge of Result mit mittlerer $\Delta = 2.42$; $p < .05$; $d = 0.54$).

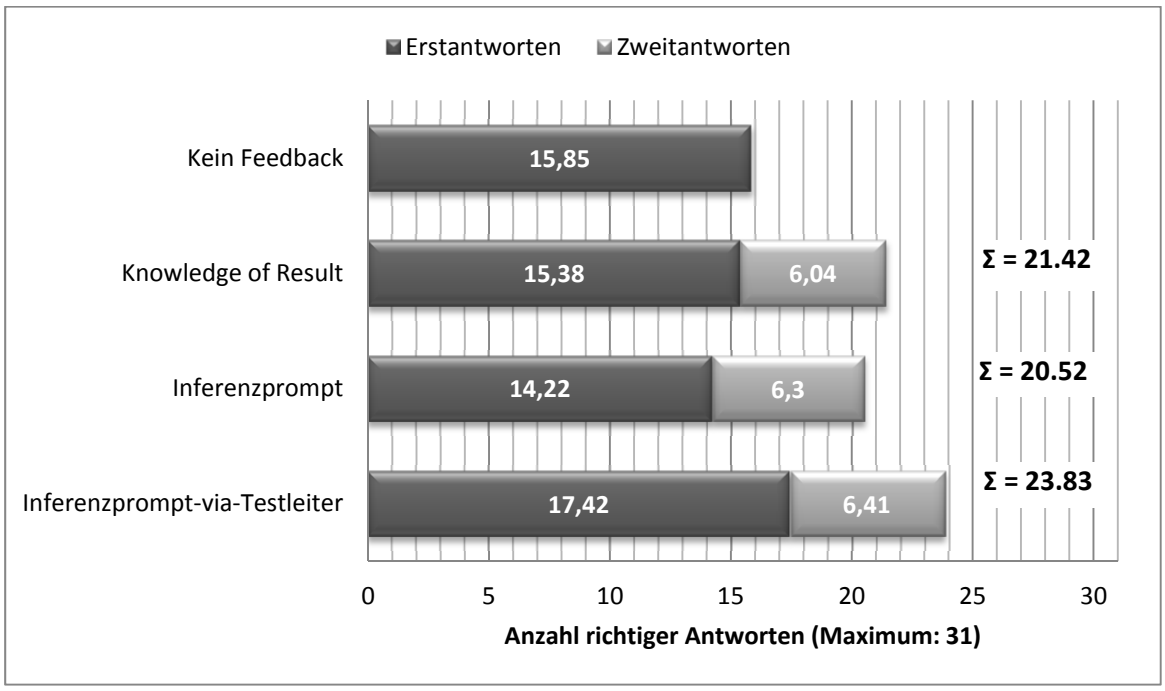

Abbildung 16 Gemeinsame Darstellung der durchschnittlichen Anzahl richtiger Antworten in Erst- und Zweitversuchen (N = 230).

11.3.2 Die Leistung im Posttest

Neben den berichteten Auswirkungen der Feedbackinterventionen auf die Leistung in der Treatmentphase ist die Wirksamkeit der Feedbackinterventionen durch ihre Auswirkungen im Posttest zu beurteilen. Die deskriptive Statistik hierfür ist in Tabelle 42 wiedergegeben. Die ANOVA belegt einen signifikanten Haupteffekt der

Feedbackinterventionen auf die Leistung im Posttest (F (3, 205) = 4.70; p < .01; η^2 = .06).

Die Post-hoc-Tests (Scheffé) zeigen, dass sich die signifikanten Leistungsunterschiede aus dem Kontrast der Bedingung Inferenzprompt-via-Testleiter zur Gruppe Inferenzprompt (mittlere Differenz Δ = 1.95; p < .05) einerseits und zur feedbackfreien Kontrollgruppe (mittlere Δ = 1.73, p < .05) andererseits ergeben: Die Testleiterbedingung bringt mit einer durchschnittlichen Leistung von M = 8.80 die höchste Leistung hervor und konnte damit im Mittel M = 1.72 Aufgaben des Posttests mehr korrekt beantworten als die Bedingung Inferenzprompt. Dieser Mittelwertunterschied drückt sich in einer mittleren Effektstärke von d = 0.57 (Cohens d) aus. In Relation zur feedbackfreien Kontrollgruppe hatte die Testleiterbedingung im Durchschnitt annähernd zwei Aufgaben mehr lösen können, was einer Effektstärke von Cohens d = 0.67 entspricht.

Die Gruppe Knowledge of Result ist mit ihrer Leistung am nächsten an der Bedingung Inferenzprompt-via-Testleiter. Die Leistungen beider Gruppen weichen nicht statistisch signifikant voneinander ab (p > .05).

Tabelle 42 Deskriptive Statistik der Leistung im Posttest (N = 209)

		Leistung im Posttest (14 Items)		
	N	*M*	*SD*	*M* %
Kein Feedback	49	6.86	3.12	49.00
Knowledge of Result	53	7.77	2.72	55.50
Inferenzprompt	51	7.08	3.30	50.57
Inferenzprompt-via-Testleiter	56	8.80	2.67	62.86

11.3.3 Vergleich der Leistungen in der Treatment- und der Posttestphase

Die Betrachtung der durchschnittlichen Posttestleistung der Versuchsgruppen in Prozent (vgl. Tabelle 42) lenkt die Aufmerksamkeit auf die entsprechende prozentuale Testleistung aus der Treatmentphase und die Frage, ob sich beide Tests darin unterscheiden, wie viele ihrer Aufgaben im Durchschnitt von den Versuchsgruppen korrekt gelöst werden konnten und ob sich die Versuchsgruppen hierin unterscheiden. Zur Überprüfung wurde eine ANOVA mit Messwiederholung (RM-ANOVA) durchgeführt.

Der Messzeitpunkt (Treatment/Posttest) stellt den Innersubjektfaktor dar, die Versuchsgruppen den Zwischensubjektfaktor.

Für diese Analyse wurden nur die Fälle berücksichtigt, die beide Tests abgeschlossen und die dafür jeweils festgelegten Kriterien bezüglich „durchgeklickter“ Items (vgl. 10.8.1) erfüllten. Die Auswertungsstichprobe reduzierte sich deshalb auf N = 202 Fälle. Darüber hinaus wurden hier relative Lösungshäufigkeiten berücksichtigt, da die Testteile der Treatmentphase und des Posttests unterschiedlich viele Items beinhalten. Die entsprechenden deskriptiven Kennwerte sind in Tabelle 43 dargestellt (vgl. auch Abbildung 17). Daneben sind noch die mittleren Differenzwerte aus beiden Testteilen angegeben. Ein negativer Differenzwert, wie er in allen drei Feedbackbedingungen entsteht, bedeutet, dass im Posttest relativ mehr Aufgaben gelöst wurden als in den ersten Versuchen des Tests der Treatmentphase.

Tabelle 43 Deskriptive Statistik der relativen Lösungshäufigkeiten im Treatment und im Posttest sowie der Differenz daraus (N = 202)

		Treatment (Erstantworten)		Posttest		Differenzwert (Treatment-Posttest)	
	N	*M*	*SD*	*M*	*SD*	*M*	*SD*
Kein Feedback	46	0.51	0.15	0.51	0.22	0.01	0.18
Knowledge of Result	51	0.50	0.16	0.56	0.19	-0.07	0.16
Inferenzprompt	49	0.48	0.19	0.51	0.24	-0.03	0.18
Inferenzprompt-via-Testleiter	56	0.56	0.15	0.63	0.19	-0.07	0.14
Gesamt	202	0.51	0.16	0.56	0.22	-0.04	0.17

Aus der RM-ANOVA resultiert zum einen ein signifikanter Effekt des Messzeitpunktes (F (1, 198) = 11.88; $p < .01$; partielles $\eta^2 = .06$): im Posttest fällt die Testleistung um 0.04 Einheiten bzw. 4 % höher aus als in der Treatmentphase ($p < .01$). Zum anderen erweisen sich die Unterschiede zwischen den Versuchsgruppen ebenfalls als statistisch signifikant (F (3, 198) = 4.02; $p < .05$; partielles $\eta^2 = .06$). Der signifikante Unterschied ergibt sich hier, den Scheffé-Tests zufolge, aus der höheren Leistung der Gruppe Inferenzprompt-via-Testleiter gegenüber der Bedingung Inferenzprompt. Darüber hinaus erbringt die Interaktion beider Faktoren keinen statistisch signifikanten Effekt (F (3, 198) = 2.21; $p > .05$; partielles $\eta^2 = .03$).

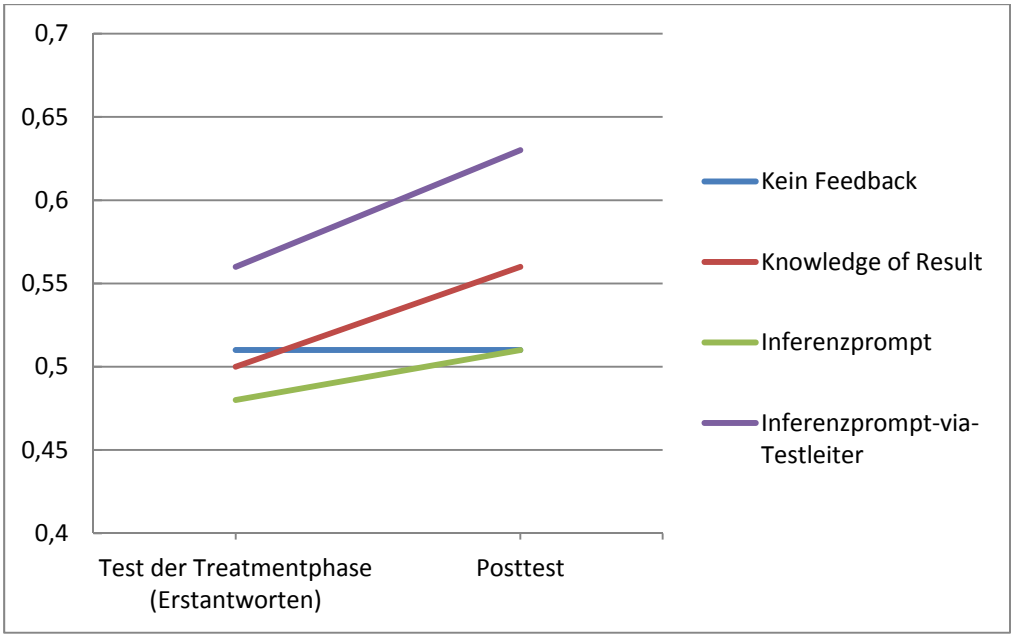

Abbildung 17 Relativierte Testleistung in der Treatmentphase und dem Posttest.

11.4 Haupteffekt von Feedback auf die Bearbeitungszeiten

Zunächst sind die Daten der Bearbeitungszeiten hinsichtlich möglicher Ausreißerwerte analysiert worden. Als Ausreißerwerte sind hier Werte definiert, die das Anderthalbfache des Interquartilsabstandes (IQA) unterhalb des ersten Quartils (Q1) oder oberhalb des dritten Quartils (Q3) liegen (Q3 + 1.5*IQA < x < Q1 - 1.5*IQA). Ausreißerwerte wären, um die deskriptiven Kennwerte nicht zu verzerren, aus den Analysen der Bearbeitungszeiten auszuschließen (vgl. Abschnitt 6.4). Für dieses Experiment liegen jedoch keine Ausreißerwerte vor; alle Fälle konnten berücksichtigt werden.

Treatmentphase

Zur Beurteilung der Bearbeitungszeiten wurde zunächst die Zeit analysiert, die für die gesamte Treatmentphase, das heißt für alle 31 Items und in den Feedbackbedingungen inklusive der zweiten Versuchen, aufgewendet wurde. Die Mittelwerte und Standardabweichungen sind in Tabelle 44 zusammengefasst. Die ANOVA zeigt hier einen signifikanten Haupteffekt der Feedbackinterventionen auf die Bearbeitungszeit (F (3, 226) = 28.24; p < .001; η^2 = .27).

Dieser Effekt geht, den Post-hoc-Tests (Scheffé) zufolge, auf die Bedingung Inferenzprompt-via-Testleiter zurück, die für den Abschluss der Experimentalphase ungefähr zehn Minuten mehr als jede andere Bedingung benötigte (jeweils $p < .001$). Die Höhe dieser Unterschiede wird durch große Effektstärken von Cohens $d = 1.30$ (Kontrollgruppe) bzw. $d = 1.31$ (Knowledge of Result) sowie $d = 1.39$ (Inferenzprompt) reflektiert. Die anderen Versuchsgruppen unterscheiden sich nicht signifikant voneinander.

Tabelle 44 Deskriptive Statistik der Bearbeitungszeiten in der Treatmentphase (N = 230)

		Aufgewendete Minuten					
		Gesamt		Erstantworten (31 Items)		Zweitantworten	
	N	*M*	*SD*	*M*	*SD*	*M*	*SD*
Kein Feedback	53	29.82	7.32	29.82	7.32	–	–
Knowledge of Result	58	29.84	7.16	27.08	6.45	2.76	1.64
Inferenzprompt	60	29.48	6.80	25.89	6.41	3.59	1.78
Inferenzprompt-via-Testleiter	59	40.01	8.32	31.78	6.99	8.23	2.90

Anmerkungen. – nicht verfügbar (Kontrollgruppe ohne Zweitversuche).

In einem weiteren Schritt wurde die Gesamtbearbeitungszeit aufgeteilt auf die Anteile, die auf die Erstantworten einerseits und die Zweitantworten andererseits entfallen (vgl. Tabelle 44 für deskriptive Statistiken). Die varianzanalytische Auswertung der Zeit, die nur für die ersten Antwortversuche aufgebracht wurde, ergibt erwartungsgemäß einen signifikanten Haupteffekt (F (3, 226) = 9.04; $p < .001$; $\eta^2 = .11$). Dabei belegen die Post-hoc-Tests (Scheffé), dass sich die beiden Bedingungen Knowledge of Result und Inferenzprompt auch hier nicht signifikant voneinander unterscheiden ($p > .05$). Daneben heben sie sich wiederum signifikant von der Testleiterbedingung ab, indem die Gruppe Knowledge of Result durchschnittlich knapp fünf Minuten ($p < .01$) und die Gruppe Inferenzprompt im Mittel fast sechs Minuten ($p < .001$) weniger Zeit auf die Erstantworten verwendete. Außerdem arbeitete die Bedingung Inferenzprompt signifikant kürzer an den Erstantworten als die Kontrollgruppe (mittlere $\Delta = - 3.92$; $p < .05$). Darüber hinaus liegt kein signifikanter Unterschied zwischen der Kontrollgruppe und der Bedingung Inferenzprompt-via-Testleiter vor (mittlere $\Delta = |1.97|$; $p > .05$; $d = 0.27$).

Zur Prüfung von Gruppenunterschieden hinsichtlich der Zeit, die auf alle zweiten Versuche entfällt, wird aufgrund inhomogener Fehlervarianzen (Levene-Test; $F\,(2, 174) = 13.09$, $p < .001$) auf den Kruskal-Wallis-Test zurückgegriffen. Die Analyse weist hier auf signifikante Unterschiede hin ($\chi^2 = 104.49$; $p < .001$). Dabei wiederholt sich das Bild aus den Analysen der Gesamtbearbeitungszeit und der Zeit für die Erstantworten: dem Scheffé-Test zufolge ist der Unterschied zwischen Knowledge of Result und Inferenzprompt mit einer mittleren Differenz $\Delta = |0.83|$ ($p > .05$; $d = |0.49|$) nicht statistisch bedeutsam. Die Gruppe Inferenzprompt-via-Testleiter wendete im Durchschnitt mehr als doppelt so viel Zeit für die Zweitantworten auf als die beiden anderen Gruppen und der Scheffé-Test belegt die statistische Bedeutsamkeit dieser Abweichungen gegenüber der Gruppe Knowledge of Result (mittlere $\Delta = 5.47$; $p < .001$; Cohens $d = 2.32$) und Inferenzprompt (mittlere $\Delta = 4.64$; $p < .001$; Cohens $d = 1.93$). Da es sich hier um die absolute Anzahl der Minuten, die für alle zweiten Versuche aufgebracht wurde, handelt, ist zu berücksichtigen, dass die Testleiterbedingung im Durchschnitt auch weniger zweite Versuche benötigte als die beiden anderen Feedbackgruppen.

Da die Gesamtzeit für die Zweitantworten auch von der Anzahl benötigter zweiter Versuche abhängt, ist für einen Vergleich der Dauer der Auseinandersetzung mit einer Erst- und einer Zweitantwort die Zeit zu vergleichen, die *durchschnittlich* für *eine* Erstantwort und im Vergleich dazu für eine Zweitantwort aufgewendet wurde. Dabei ist für die Erstantworten allerdings zu berücksichtigen, dass mit der ersten Präsentation eines neuen Textes gleichzeitig auch das erste Item präsentiert wird. Da bei der Bearbeitung der Units eher erwartet werden kann, dass zunächst der Text vollständig gelesen wird und dann die Items beantwortet werden, fällt die Bearbeitungszeit für das erste Item sehr wahrscheinlich mit der Textrezeption zusammen, wodurch für das erste Item einer jeden Unit vermutlich verhältnismäßig mehr Zeit als für die ersten Versuche der restlichen Items aufgewendet werden. Die entsprechenden Analysen des ersten Experiments bestätigten diese Vermutung (vgl. Abschnitt 6.4, Tabelle 27).

Die Analyse der durchschnittlichen Zeit für eine Erstantwort, und zwar bezogen auf erste Items der Units einerseits und alle weiteren Items andererseits, wurde als RM-ANOVA durchgeführt. Der Innensubjektfaktor ist der Messzeitpunkt (also erste Items einer Unit vs. restliche Items einer Unit) und der Zwischensubjektfaktor ist die

Versuchsgruppenzugehörigkeit. Die deskriptiven Kennwerte sind in Tabelle 45 zusammengefasst. Die RM-ANOVA belegt einen hoch signifikanten Effekt des Innersubjektfaktors (F (1, 226) = 1051.46; $p < .001$; partielles $\eta^2 = .82$), ebenso des Zwischensubjektfaktors (F (1, 226) = 8.14; $p < .001$; partielles $\eta^2 = .10$) und auch die Interaktion erweist sich als signifikant (F (1, 226) = 6.86; $p < .001$; partielles $\eta^2 = .08$). Das bedeutet, die Erstantworten für erste Items einer Unit gehen mit sehr viel längerer Bearbeitungszeit einher als für die restlichen Items. Somit wurden die Texte im Allgemeinen offenbar am Anfang einer neuen Unit zunächst (vollständig oder zumindest ausführlicher) gelesen, bevor die Aufgaben beantwortet wurden. Der signifikante Effekt der Versuchsgruppen geht wiederum auf die Gruppe Inferenzprompt zurück, die signifikant weniger Zeit für Erstantworten aufgewendet hat als die Kontrollgruppe (mittlere $\Delta = -20.61$; $p < .001$) und die Testleiterbedingung (mittlere $\Delta = -28.40$; $p < .001$).

Analog zu den beiden Aspekten der Erstantworten lässt sich für die Feedbackbedingungen auch die entsprechenden Zeiten für eine durchschnittliche Zweitantwort mit der Zeit für eine Erstantwort (ohne erste Items einer Unit) vergleichen. Die RM-ANOVA belegt hier einen hoch signifikanten Effekt für den Innersubjektfaktor, das bedeutet den Vergleich zwischen der durchschnittlich aufgewendeten Zeit für eine Erst- versus eine Zweitantwort (F (1, 174) = 285.43; $p < .001$; partielles $\eta^2 = .62$). Der Zwischensubjektfaktor ist ebenfalls signifikant (F (2, 174) = 84.99; $p < .001$; partielles $\eta^2 = .50$) und entsprechend ist auch die Interaktion signifikant (F (2, 174) = 86.28; $p < .001$; partielles $\eta^2 = .50$). Das bedeutet, für Zweitantworten wird im Allgemeinen signifikant weniger Zeit aufgewendet als für eine Erstantwort. Der signifikante Effekt der Versuchsgruppen geht wiederum auf die Testleiterbedingung zurück, die den Post-hoc-Tests (Scheffé) zufolge signifikant mehr Zeit auf die Beantworten von Items aufgewendet hat als die Gruppe Inferenzprompt (mittlere $\Delta = 13.73$; $p < .001$) oder Knowledge of Result (mittlere $\Delta = 16.45$; $p < .001$).

Tabelle 45 Bearbeitungszeiten auf Itemebene (N = 230)

		Zeit für eine durchschnittliche Antwort (in Sekunden)					
		Erstantwort				Zweitantwort	
		1. Items der Units (eher Textrezeption)		Alle Items ohne die 1. Items pro Unit			
	N	*M*	*SD*	*M*	*SD*	*M*	*SD*
Kein Feedback	53	178.11	59.65	36.01	9.93	–	–
Knowledge of Result	58	169.30	63.70	31.76	9.74	11.45	6.95
Inferenzprompt	60	138.76	61.40	34.13	8.58	14.51	8.24
Inferenzprompt-via-Testleiter	59	192.26	67.86	37.41	7.37	38.69	11.76

Anmerkungen. – nicht verfügbar (Kontrollgruppe ohne Zweitversuche).

Die Korrelation der Bearbeitungszeiten von Erst- und Zweitantworten (hier Gesamtwert, nicht auf Itemebene) ist mit $r = .40$ ($p < .001$) statistisch bedeutsam. Zusammenhänge zwischen der Bearbeitungsdauer für Erst- bzw. Zweitantworten und den entsprechenden Leistungen sind nicht nachweisbar. Die Korrelation der Bearbeitungszeit für Erstantworten und die Leistung in den Erstantworten ergibt $r = .09$ ($p > .05$). Ebenso ist zwischen der Bearbeitungszeit für die Zweitantworten und die relative Lösungshäufigkeit in den Zweitversuchen mit $r = .09$ ($p > .05$) kein Zusammenhang nachweisbar.

Posttest

Die Versuchsgruppen unterscheiden sich auch in der Zeit, die für den Posttest aufgewendet wurde (vgl. Tabelle 47). Die ANOVA belegt hierfür einen statistisch signifikanten Haupteffekt der Feedbackintervention (F (3, 205) = 3.29; $p < .05$; $\eta^2 = .05$). Dieser Effekt geht, wie die Post-hoc-Tests (Scheffé) zeigen, allerdings ausschließlich auf den Unterschied der beiden Feedbackbedingungen Inferenzprompts und Inferenzprompts-via-Testleiter zurück, wobei die Testleiterbedingung die meiste Zeit für den Posttest aufgewendet hat. Die mittlere Differenz beider Gruppen beträgt $\Delta = 1.54$ ($p < .05$) und präsentiert sich mit einer Effektstärke von Cohens $d = 0.55$ als mittlergroßer Effekt. Die restlichen Gruppenkontraste (Scheffé-Test) erreichen keine statistische Signifikanz.

Tabelle 46 Deskriptive Statistik der Bearbeitungsdauer des Posttests (N = 209)

	Aufgewendete Minuten in Posttest (14 Items)		
	N	*M*	*SD*
Kein Feedback	49	9.29	2.55
Knowledge of Result	53	8.74	2.57
Inferenzprompt	51	8.19	2.95
Inferenzprompt-via-Testleiter	56	9.73	2.67

11.5 Analyse der Testmotivation

Die deskriptive Statistik zu den Angaben der Probanden, wie sehr sich bei der Bearbeitung des Tests (Treatment) angestrengt hatten, ist in Tabelle 47 wiedergegeben. Dabei fällt auf, dass sich die Kennwerte zwischen den Gruppen stark ähneln. Daneben sprechen die Lageparameter dafür, dass die Untersuchungsteilnehmer im Allgemeinen angaben, sich eher sehr angestrengt zu haben.

Tabelle 47 Deskriptive Statistik zur Testmotivation (N = 228)

	Testmotivation (Einzelindikator)				
	N	*M*	*SD*	*Med*	*Mo*
Kein Feedback	53	7.32	2.25	8.00	9.00
Knowledge of Result	57	7.82	1.59	8.00	8.00
Inferenzprompt	59	7.54	1.87	8.00	8.00
Inferenzprompt-via-Testleiter	59	7.75	1.90	8.00	9.00

Anmerkungen. Die Skala reicht von 1 für minimale Anstrengung bis 10 für maximale Anstrengung.

Bei der Überprüfung von Gruppenunterschieden hinsichtlich der Testmotivation erbrachte der Levene-Test zunächst, dass die Fehlervarianzen innerhalb der Versuchsgruppen nicht homogen sind (F (3, 224); $p < .05$). Demzufolge wurde der Kruskal-Wallis-Test verwendet. Die Analyse bestätigt, dass sich die Versuchsgruppen in ihren Angaben zur Testmotivation nicht unterscheiden ($\chi^2 = 1.30$; $p > .73$).

Zusammenhänge der Angaben zur Testmotivation mit den tatsächlichen Leistungen in der Treatmentphase sind nicht nachweisbar. Die Korrelation der Testmotivation mit der

Leistung in den Erstantworten beträgt $r = .12$ ($p > .05$) und die Korrelation zur Leistung in den zweiten Versuchen ergibt $r = .07$ ($p > .05$).

11.6 Analyse der wahrgenommenen Nützlichkeit der Feedbacks

Die deskriptiven Kennwerte, die sich aus der Einschätzung der Feedbackbedingungen hinsichtlich der der Nützlichkeit der in der Treatmentphase dargebotenen Rückmeldungen ergeben, sind in Tabelle 48 wiedergegeben. Die Kennwerte zeigen, dass sich die beiden Bedingungen Knowledge of Result und Inferenzprompt in ihren Einschätzungen sehr ähnlich sind. Die Einschätzung der Gruppe Inferenzprompt-via-Testleiter fällt dagegen im Mittel etwas höher aus.

Zur Prüfung möglicher Gruppenunterschiede wurde aufgrund inhomogener Fehlervarianzen (Levene; F (2, 172) = 4.47, $p < .05$) auf den Kruskal-Wallis-Test zurückgegriffen. Der Kruskal-Wallis-Test belegt einen signifikanten Unterschied zwischen den Gruppen ($\chi^2 = 13.96$, $p < .01$). Scheffé-Tests zeigen, dass die Gruppen Inferenzprompt und Knowledge of Result sich mit einer mittleren Differenz von $\Delta = |0.05|$ ($p > .05$) nicht signifikant voneinander abheben. Im Vergleich zur Testleiterbedingung fallen die Einschätzungen der beiden Gruppen jedoch signifikant geringer aus: die mittlere Differenz der Gruppe Knowledge of Result gegenüber der Testleiterbedingung beträgt $\Delta = - 0.37$ ($p < .05$). Die mittlere Abweichung der Gruppe Inferenzprompt gegenüber der Testleiterbedingung liegt im Mittel bei $\Delta = - 0.41$ ($p < .01$).

Tabelle 48 Deskriptive Statistik zu Einschätzungen der Feedbacks (N = 175)

		Nützlichkeit der Feedbacks (5 Items)					
	N	*M*	*SD*	*Med*	*Mo*	*Min*	*Max*
Knowledge of Result	57	2.74	0.71	2.80	2.80; 3.20	1.00	4.00
Inferenzprompt	59	2.70	0.66	2.80	3.00	1.20	3.80
Inferenzprompt-via-Testleiter	59	3.11	0.52	3.20	3.20	1.40	4.00

Anmerkungen. Die Beantwortung der Items erfolgte auf einer Skala von 1 für „stimmt gar nicht“ bis 4 für „stimmt völlig“.

Über die Gruppenunterschiede in der Einschätzung der Nützlichkeit der Rückmeldungen hinaus liegt jedoch kein signifikanter Zusammenhang mit der Leistung vor. Für die Leistung in den Erstantworten ergibt sich ein Korrelation von r = .02 (p > .05), für die Leistung in den Zweitantworten ein Koeffizient von r = .004 (p > .05).

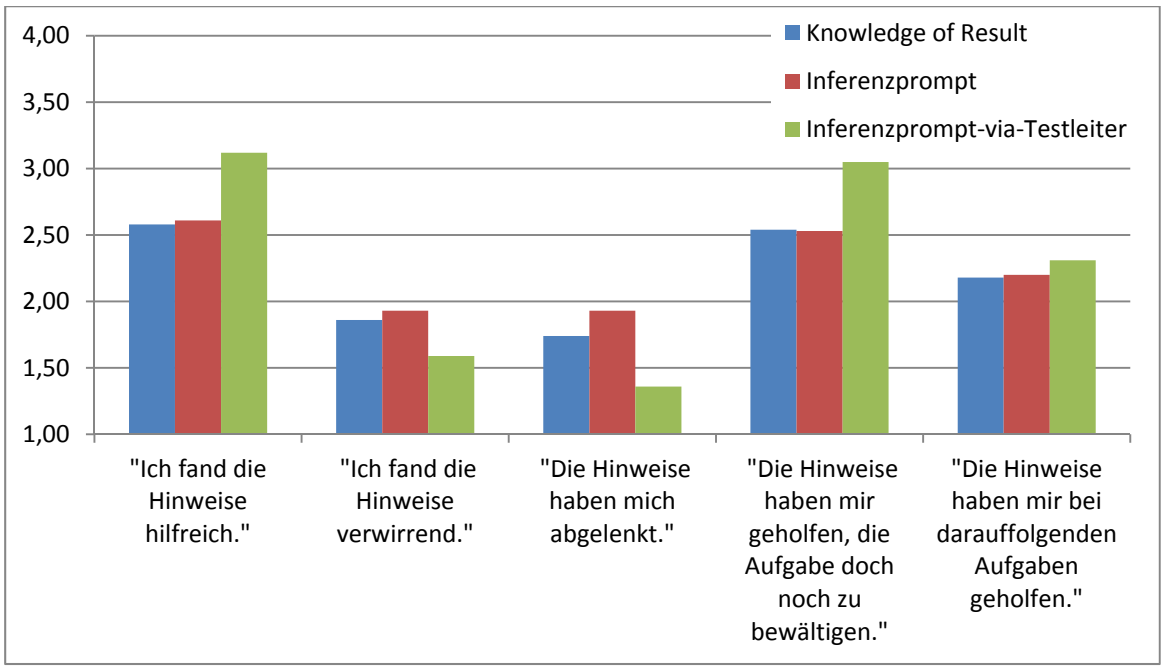

Abbildung 18 Gruppenmittelwerte (Einzelitems) in Einschätzung der Nützlichkeit der Rückmeldungen.

Anmerkungen. Skalierung von 1 für „stimmt gar nicht“ bis 4 für „stimmt völlig“, die abgebildeten Mittelwerte von Item 2 und Item 3 (von links) basieren auf den nicht rekodierten Rohwerten.

11.7 Zusammenfassung der zentralen Ergebnisse

Die für die Beantwortung der Fragestellungen zentralen Ergebnisse können wie folgt zusammengefasst werden. Bezüglich der *Auswirkungen auf das Textverständnis bzw. die Lesekompetenz* zeigt sich, dass die Bedingung Inferenzprompts-via-Testleiter:

- in den Zweitantworten eine signifikant höhere Leistung erbrachte als die Bedingungen Inferenzprompts und Knowledge of Result,
- in den Erstantworten signifikant besser abschnitt als die Bedingung Inferenzprompts,

- im Posttest eine signifikant höhere Leistung erbrachte als die Bedingung Inferenzprompt und die Kontrollgruppe ohne Feedback.

Darüber hinaus waren keine signifikanten Leistungsunterschiede zwischen den Versuchsgruppen zu finden. Damit wurden die a-priori aufgestellten Hypothesen bezüglich der Auswirkungen der Feedbackinterventionen auf die Leistung teilweise bestätigt. Entgegen der Hypothesen erweist sich, dass die Bedingung Inferenzprompt-via-Testleiter in den Erstantworten gleich auf mit der Bedingung Knowledge of Result und der feedbackfreien Kontrollgruppe lag sowie im Posttest keinen signifikanten Leistungsvorteil gegenüber der Bedingung Knowledge of Result erbracht hat.

Aus den Analysen der *Bearbeitungszeiten* ist zusammenfassend festzuhalten, dass die Bedingung Inferenzprompt-via-Testleiter:

- für die Erstantworten und für eine durchschnittliche Zweitantwort signifikant mehr Zeit für die Bearbeitung aufwendete als die Bedingungen Knowledge of Result und Inferenzprompt,
- für den Posttest signifikant mehr Zeit aufbrachte als die Bedingung Inferenzprompts,
- im Vergleich zur feedbackfreien Kontrollgruppe weder länger an den Erstantworten noch am Posttest arbeitete.

Fernen belegen die Analysen, dass:

- die Angaben zur *Anstrengungsmotivation* im Allgemeinen hoch ausgefallen sind,
- hinsichtlich der Anstrengungsmotivation keine Gruppenunterschiede und auch keine Zusammenhänge zur erbrachten Leistung im Treatment nachweisbar waren,
- die *Nützlichkeit der Rückmeldungen* im Allgemeinen als eher positiv eingestuft wurde,
- die Einschätzungen hinsichtlich der Nützlichkeit der Rückmeldungen in der Bedingung Inferenzprompts-via-Testleiter signifikant höher ausfielen als in den beiden anderen Gruppen, die sich hierin zudem nicht bedeutsam voneinander unterscheiden,
- die Einschätzungen zur Nützlichkeit der Rückmeldungen keine signifikanten Zusammenhänge zu den erbrachten Leistungen aufweisen.

12 Diskussion

Die Rolle des Experiments im Rahmen dieser Arbeit bestand darin, ausgehend von der Ineffektivität der elaborierten Feedbackinterventionen im ersten Experiment, dem Nutzen der Inferenzprompts in einem veränderten Setting nachzugehen. Das Neue an dem Setting war, dass das elaborierte Feedback nicht über den Computer, sondern (mündlich) über den Testleiter und in einer Einzelsitzung gegeben wurde (vgl. Abschnitt 8 für Begründung). Dabei wird angenommen, dass mittels der persönlichen, testleitergebundenen Präsentation auf Seiten der Probanden eine höhere Verbindlichkeit gegenüber der Umsetzung der Feedbackmitteilungen erzeugt werden kann und dadurch die Voraussetzungen für die potentielle Wirkung des Feedbacks verbessert werden. Somit ist in dem vorliegenden Experiment neben dem Feedbackinhalt auch die Art seiner Präsentation berücksichtigt worden. Die zentrale Fragestellung war weiterhin auf die Wirksamkeit der elaborierten Feedbackintervention auf das Textverständnis/die Lesekompetenz gerichtet.

Folgende Feedbackbedingungen wurden neben Inferenzprompts-via-Testleiter umgesetzt: die einfachste Art der Rückmeldung, Knowledge of Result, und die Inferenzprompts, beide als ausschließlich computervermittelte Feedbackvarianten. Als vierte Versuchsgruppe wurde die Testbedingung ohne Feedback hinzugenommen. Im Grunde stellen alle drei zuletzt genannten Gruppen Kontrollbedingungen für die Gruppe Inferenzprompts-via-Testleiter dar. Die Kontrollgruppe ohne Feedback bildet die Leistungsfähigkeit der neuen Stichprobe unter einer herkömmlichen Testbedingung an. Die Feedbackbedingung Knowledge of Result ist insofern wieder eine Kontrollbedingung, weil sie nachweislich keinen Effekt auf das Textverstehen hat, dabei aber durch die identische Feedbackprozedur einen faireren Vergleich zu der interessierenden, elaborierten Feedbackintervention bietet. Die Inferenzprompts als rein computervermittelte Variante werden ebenfalls wiederholt, obwohl sie im ersten Experiment keinen Effekt auf die Leistung ausübten. Aber durch das unvollständige, zweifaktorielle Design (Feedbackinhalt, Präsentationsmodus) wären die Vergleichbarkeit bzw. die Interpretationsmöglichkeiten der Ergebnisse ohne diese Bedingung nicht gut möglich (vgl. Abschnitt 8).

Der Fokus dieses Experimentes lag also auf der Feedbackbedingung Inferenzprompts-via-Testleiter. Es wurde vermutet, dass sie im Vergleich zu den Bedingungen Knowledge of Result und Inferenzprompts sowie der feedbackfreien Kontrollgruppe zu signifikanten Leistungssteigerungen in den Erst- und Zweitantworten der Treatmentphase sowie im Posttest führt. Von den Bedingungen Knowledge of Result und Inferenzprompts wurde erwartet, dass sie keinen Effekt auf das Textverstehen/die Lesekompetenz ausüben und sich deshalb nicht signifikant von der Leistung der feedbackfreien Kontrollgruppe abheben.

12.1 Über die Auswirkungen der Feedbackinterventionen auf das Textverständnis/die Lesekompetenz

Die Erwartungen an die testleitergebundene Feedbackgabe von Inferenzprompts in Bezug auf die Verbesserung des Textverstehens sind in Teilen bestätigt worden. Die Resultate der anderen Gruppen bestätigen deren Ineffektivität, wie sie sich im ersten Experiment zeigte. Die Überlegungen zu den Gründen ihrer Wirkungslosigkeit werden hier nicht wiederholt, sondern erst in der abschließenden Diskussion beider Experiment (vgl. Abschnitt 12.5) wieder aufgegriffen. Die nachfolgende Diskussion der Auswirkungen der Feedbackbedingungen auf die Leistung wird stattdessen, gemäß dem Anliegen dieses Experiments, aus der Perspektive der Bedingung Inferenzprompts-via-Testleiter geführt.

Die Feedbackbedingung Inferenzprompts-via-Testleiter wirkte sich also nicht in jeder erwarteten Beziehung leistungssteigernd aus. Betroffen ist hier die Transferleistung (d.h. Erstantworten in Treatment und Posttest). Insofern ist keine pauschale Schlussfolgerung bezüglich positiver Effekte der Inferenzprompts-via-Testleiter auf die Leistung zu ziehen.

Zweitantworten (Treatmentphase)

Die Korrekturleistung falscher erster Antworten fiel in der Bedingung Inferenzprompts-via-Testleiter signifikant höher aus als in den beiden Bedingungen mit rein computervermitteltem Feedback (Knowledge of Result und Inferenzprompts), die sich in der Leistung in den Zweitantworten wiederum nicht voneinander unterschieden. Dieser Befund ist eindeutig – Inferenzprompts, sofern sie über einen Testleiter vermittelt

werden, verhelfen den Lesern dazu, eine bedeutsame Anzahl an Aufgaben, die initial falsch beantwortet wurden, doch noch zu richtig zu lösen. Das impliziert, dass Inferenzprompts eine in dieser Hinsicht effektive Feedbackart darstellen. Sie unterstützen das Ziehen von Inferenzen.
Annähernd die Hälfte der falschen Erstantworten wurde in der Testleiterbedingung nach der Feedbackpräsentation richtig beantwortet. Im Vergleich dazu hat die Bedingung Inferenzprompt und Knowledge of Result jeweils ungefähr 40 % der Fehler der ersten Versuche korrigieren können. Dieser Mehrwert von durchschnittlich annähernd 10 % kann als mittlerer Effekt bewertet werden. Unter Berücksichtigung des Charakters der Intervention (d.h. Kürze der Intervention, komplexe Anforderungen) kann dieses Ergebnis positiv unterstrichen werden.

Eine mögliche Erklärung, warum diese Leistungssteigerung mithilfe der Inferenzprompts möglich wird, wenn sie über den Testleiter vermittelt werden, bieten die Bearbeitungszeiten. Auffällig ist, dass die Feedbackbedingung Inferenzprompts-via-Testleiter sehr viel mehr Zeit für eine Zweitantwort aufwendete als die beiden anderen Feedbackbedingungen. Eine durchschnittliche Zweitantwort in der Testleiterbedingung nahm etwa 39 Sekunden in Anspruch, im Vergleich dazu wurden in den beiden computervermittelten Feedbackbedingungen annähernd 12 Sekunden (Knowledge of Result) bzw. 15 Sekunden (Inferenzprompts) aufgewendet. Selbst wenn der testleitervermittelten Feedbackgabe zugestanden wird, dass die mündliche Übermittlung der Information etwas mehr Zeit in Anspruch nehmen könnte als die computervermittelte Präsentation, würde das dennoch nicht den sehr deutlichen Unterschied der Bearbeitungszeit erklären.
Natürlich ist auch hier wie an anderer Stelle Vorsicht geboten, aus der Bearbeitungszeit auf die Handlungen oder Kognitionen zu schließen. Aber es erscheint die Aussage zulässig, dass durch die testleitergebundene Feedbackgabe die Dauer der Auseinandersetzung mit einer Aufgabe infolge einer Feedbackgabe angehoben wurde und die Probanden dabei (deshalb) erfolgreich waren.

Erstantworten (Treatmentphase)

In den Erstantworten konnte die Bedingung Inferenzprompts-via-Testleiter eine signifikante Leistungssteigerung gegenüber der Bedingung Inferenzprompts erreichen,

nicht jedoch gegenüber der Bedingung Knowledge of Result und der feedbackfreien Bedingung. Dieser Befund ist insofern nicht einfach zu beurteilen, da der Mehrwert der elaborierten Feedbackintervention eigentlich (auch) gegenüber der Bedingung mit dem einfachen Feedback Knowledge of Result und/oder gegenüber der feedbackfreien Testbedingung zu bemessen ist. Denn, wie bereits an verschiedenen Stellen erläutert, ist davon auszugehen, dass Knowledge of Result keinen Effekt auf das Textverstehen hat, dabei aber unter denselben Bedingung der Feedbackprozedur (mit Mehrarbeit durch Zweitantworten, Unterbrechungen durch Feedback, etc.) gearbeitet hat und somit den faireren Vergleich zur Beurteilung eines Effekts einer elaborierten Feedbackintervention (hier Inferenzprompts-via-Testleiter) bietet. Insofern ist zu hinterfragen, worauf der gefundene signifikante Leistungsunterschied zwischen den beiden Bedingungen mit Inferenzprompts basiert und welche Schlüsse hinsichtlich des Effekts zugunsten der Testleiterbedingung daraus zu ziehen sind.

Dass allein der Leistungsunterschied zwischen den beiden Gruppen mit Inferenzprompts signifikant ist, wäre damit zu erklären, dass erstens die Bedingung der rein computervermittelten Inferenzprompts zu einer etwas (aber nicht bedeutsam) schlechteren Leistung tendiert als die Bedingung Knowledge of Result, die ihrerseits wirkungslos, weil auf dem Niveau der feedbackfreien Gruppe liegend, ist. Eine Erklärung hierfür kann sein, dass die Inferenzprompts aufgrund ihrer längeren und vor allem kognitiv fordernden Inhalte Anforderungen an den Leser stellen, die er in der anonymen Bearbeitungssituation der computerbasierten Feedbackbedingung nicht nur nicht erfolgreich bewältigt, sondern die ihn vermittelt über negative (motivationale) Prozesse (z.B. Frustration) ungenauer oder unbedachter antworten lassen. Durch das tendenzielle Abschwächen der Leistung der Bedingung Inferenzprompts erreicht dann zweitens der Vorteil der Bedingung Inferenzprompts-via-Testleiter erst ihre Signifikanz.

Aber wie ist der Befund zugunsten der Testleiterbedingung vor dem Hintergrund dieser Überlegungen zu bewerten? Sind Inferenzprompts-via-Testleiter nützlich, um die Leistung von Probanden dahingehend anzuheben, dass sie im Verlauf des feedbackgestützten Testes schon einen Transfer durchführen, also mehr Aufgaben bereits im ersten Versuch richtig beantworten? Die Untersuchung bietet verschiedene Anhaltspunkte, die für die Bedeutung des Effekts sprechen und ihn unterstreichen. Zum einen ist zu berücksichtigen, dass die Leistung in den Erstantworten eine Transferleistung darstellt, die, wie an mehreren Stellen schon erläutert und nachgewiesen wurde, offensichtlich eine kognitiv anspruchsvolle, voraussetzungsvolle Anforderung darstellt.

Außerdem ist zu beachten, dass die Testbedingung, in der die Gruppe Inferenzprompts-via-Testleiter die Leistungssteigerung bewirkte, eine einmalige Sitzung darstellt. Die Treatmentphase mit den fünf Units und insgesamt 31 Items dauerte in dieser Gruppe im Durchschnitt 40 Minuten. Es ist nicht ausgeschlossen, dass der Leistungsvorteil der Inferenzprompts-via-Testleiter bei einer längeren Testphase auch gegenüber den anderen Versuchsbedingungen ausgebaut werden könnte. Denn der letztlich signifikante Leistungsvorsprung der Inferenzprompts-via-Testleiter gegenüber der Bedingung der rein computervermittelten Inferenzprompts ergibt sich erst gegen Ende des Treatments, wie die Analyse der Erstantworten über den Verlauf (über die fünf Units) der Treatmentphase zeigte.

Zum anderen ist weiterhin zu bedenken, dass die testleitergebundene Feedbackgabe, trotz des positiven Ergebnisses der Versuchsgruppe, mit den Nachteilen behaftet gewesen sein könnte, die bei der Darlegung der Forschungsbefunde zum Präsentationsmodus (vgl. Abschnitt 3.5.1.3) und im Vorfeld dieses Experiments (vgl. Abschnitt 8) angeführt wurden. Wie die Arbeiten von Kluger und Adler (1993), Comer (2007) und Guerin (1986) belegen, ist davon auszugehen, dass die Leistungsfähigkeit der Probanden aufgrund der aus ihrer Sicht vorhandenen Bewertungssituation durch den Testleiter (vgl. Abschnitt 3.5.1.3) gedämpft wurde. Die gezeigte Testleitung mit den Inferenzprompts als Hilfestellungen wäre demnach im Allgemeinen möglicherweise sogar noch hinter dem Leistungspotential zurückgeblieben.

Zwischenfazit bezüglich des Lerneffekts: Unter Berücksichtigung der Leistungen in den Erst- und den Zweitantworten wird davon ausgegangen, dass Inferenzprompts, sofern sie über den Testleiter vermittelt wurden, einen Lerneffekt hervorbrachten. In erster Linie ist hier die Korrekturleistung in den zweiten Versuchen anzuführen, die im Vergleich zu den beiden anderen Feedbackbedingungen statistisch bedeutsam höher ausfällt. Zudem konnten die Probanden dieser Bedingungen die Informationen der Rückmeldungen, offenbar eher gegen Ende der Treatmentphase, auch schon beim Beantworten der ersten Versuche nutzen. Diese Transferleistung wird nur im Vergleich zur Bedingung der computervermittelten Prompts statistisch bedeutsam, vermutlich weil in der letzteren Bedingung die Leistung tendenziell abgeschwächt wird. Das Potential der Inferenzprompts für die Transferleistung in der Treatmentphase liegt möglicherweise jedoch noch höher (auch in Bezug auf die Kontraste zu Knowledge of Result und

feedbackfreie Testbedingung), falls die Umgebungsbedingungen der testleitergebundenen Feedbackgabe auch nachteilige Effekte auf die Testleistung hatte.

Leistung im Posttest

Der Lerneffekt der Inferenzprompts-via-Testleiter offenbart sich auch im Posttest. Die Testleiterbedingung schnitt hier bedeutsam besser ab als die feedbackfreie Kontrollgruppe und wiederum die Bedingung der computervermittelten Gabe der Inferenzprompts. Die Unterschiede sind von mittlerer Stärke – die durchschnittliche Leistungsverbesserung in der Testleiterbedingung betrug jeweils etwas mehr als eine halbe Standardabweichung. Das waren in dem Posttest (14 Items) jeweils ungefähr zwei Aufgaben, die von Probanden der Testleiterbedingung mehr gelöst worden konnten.

Die Posttestleistung der Bedingung Inferenzprompts-via-Testleiter hebt sich hingegen nicht signifikant von der Leistung der Bedingung Knowledge of Result ab. Dieses Ergebnis spricht nicht gegen die Effektivität der Testleiterbedingung, schließlich besteht der Effekt gegenüber der Kontrollgruppe ohne Feedback. Vielmehr ist in Betracht zu ziehen, dass das Treatment mittels Knowledge of Result die Aufmerksamkeit der Probanden bezüglich der Testsituation angehoben hat. Damit ist gemeint, dass Knowledge of Result durchaus als eine Art Signal wirkt, das zum aufmerksameren Bearbeiten von Testaufgaben führen kann. In der Bedingung Knowledge of Result mag die Wirkung eines solchen Signals von den Probanden eher noch mit in den Posttest „mitgenommen“ worden zu sein, so dass die Posttestleistung tendenziell höher ausfiel als in der feedbackfreien Kontrollgruppe und andererseits der Leistungsvorsprung der Testleiterbedingung nicht mehr bedeutsam war. Aber, das ist noch einmal zu unterstreichen, die Bedingung Inferenzprompt-via-Testleiter schnitt am besten im Posttest ab und erbrachte dabei einen signifikanten Leistungsvorteil gegenüber der Bedingung Inferenzprompt und der feedbackfreien Kontrollgruppe.

Dieser Effekt zugunsten der Inferenzprompts-via-Testleiter wird durch die Bearbeitungszeiten des Posttests zusätzlich hervorgehoben. Denn die Testleiterbedingung arbeitete nicht länger am Posttest als die anderen Gruppen. Dieses Ergebnis ist vor dem folgenden Hintergrund ein nicht unwichtiges Argument: In der Testleiterbedingung bestand die Untersuchungssituation im Posttest noch als Einzelsetting mit anwesendem

Testleiter. Daher könnte kritisiert werden, dass der Effekt auf die Posttestleistung im Wesentlichen darauf zurückginge, dass die Bearbeitungsgenauigkeit bzw. das Testverhalten der Probanden optimal gehalten wird (Stichwort: soziale Erwünschtheit), während die Probanden in der Gruppensitzung zum Ende des Programms hin, also vor allem im Posttest, aufgrund von Ermüdung wahrscheinlich eher schneller arbeiten und dabei möglicherweise mehr unnötige Fehler produzieren.

Fazit zum Effekt der Inferenzprompts-via-Testleiter

Unter Berücksichtigung aller Ergebnisse wird die Schlussfolgerung gezogen, dass Inferenzprompts, wenn sie über einen Testleiter gegeben werden, eine effektive Unterstützung des Textverständnisses bzw. der Lesekompetenz darstellen. Sie führen dazu, dass mehr Aufgaben, die initial falsch beantwortet wurden, im zweiten Versuch richtig korrigiert werden konnten. Zudem konnten sie die Informationen der Inferenzprompts nutzen, um sie auf andere Aufgaben zu übertragen. Diese Transferleistung konnte teilweise schon gegen Ende der Treatmentphase erfolgen und entfaltete sich dann vor allem im Posttest.

12.2 Der „Testleitereffekt“

Die Resultate der Bedingung Inferenzprompt-via-Testleiter sprechen also für die Effektivität dieser Intervention. Noch nicht erörtert wurde hingegen die Frage, worin die Wirksamkeit dieser Intervention zu sehen ist. Wie trägt der Testleiter zum Effekt bei und welchen Einfluss haben die elaborierten Rückmeldungen?

Charakteristika des Settings einer testleitergebundenen Feedbackgabe

Die Variante des Testleiters als Feedbackgeber ist hauptsächlich unter der Annahme ausgewählt worden, dass dadurch eine höhere *Verbindlichkeit bzw. Anstrengungsbereitschaft* seitens der Probanden geschaffen werden kann. Durch die funktionale Einbeziehung einer Person – der Testleiter ist direkt anwesend und gibt das Feedback – ändert sich die Untersuchungssituation für den Probanden erheblich. Über dadurch wahrscheinlich stärker wirkende soziale Prozesse, etwa die soziale

Erwünschtheit betreffend, sollten die Chancen verbessert werden, dass die Probanden in der Umsetzung der Rückmeldungen nicht vermeiden.

Ebenso positiv könnte es sich in dieser Untersuchungsbedingung auswirken, dass die testleitervermittelte Feedbackgabe stärker einer *Interaktion* gleicht – eine Situation, die die Schüler aus den üblichen Lehr-Lern-Prozessen mit (Fehler-)Rückmeldungen eher gewohnt sein dürften. Damit könnte eine höhere Authentizität der Feedbackgabe einhergehen. Ebenso zeichnet sich die testleitervermittelte Feedbackpräsentation dadurch aus, dass durch die verbale Präsentation für die Probanden auch eine gewisse „Unausweichlichkeit" vor der Feedbackintervention, von Anfang bis Ende des Treatments, geschaffen wird.

Der Testleiter als Feedbackgeber impliziert im Vergleich zur computervermittelten Feedbackgabe des Weiteren eine veränderte *Modalität der Präsentation der Rückmeldungen*, die hinsichtlich ihres möglichen Einflusses auf die Wirksamkeit der Intervention zu beleuchten ist. Der Testleiter gibt das Feedback erst verbal, dann für die Dauer der Bearbeitung des zweiten Versuchs noch in schriftlicher Form. Die verbale Mitteilung stellt im Vergleich zu den anderen Feedbackbedingungen zusätzlich die Anforderung des Hörverstehens, was positive oder auch nachteilige Folgen für die Leistung haben könnte. Gegen eine Benachteiligung durch die verbale Feedbackpräsentation spricht, dass der elaborierte Teil der Rückmeldungen aus ein bis zwei Sätzen besteht und damit im Normalfall keine Überbelastung des Hörverstehens darstellen sollte. Allerdings wurde jede Rückmeldung auch in schriftlicher Form vorgelegt, so dass eventuell aufgetretene Schwierigkeiten beim Hörverstehen dadurch wettzumachen waren. Gleichzeitig schließt die zusätzlich schriftliche Präsentation eine eventuelle Überforderung der Arbeitsgedächtniskapazität aus und schafft in diesem Punkt Vergleichbarkeit zu den anderen, ausschließlich computerbasierten Feedbackbedingungen.
Denkbar ist dagegen, dass durch die verbale Feedbackübermittlung den Probanden ein Vorteil verschafft wird, wenn durch das Hören eventuell vorhandene Schwierigkeiten beim Lesen der Rückmeldung umgangen werden. Allerdings ist die verbale Feedbackpräsentation in dieser Untersuchung untrennbar mit dem Testleiter bzw. den Merkmalen des Testsettings einerseits und der zusätzlichen schriftlichen Präsentation andererseits verbunden, so dass keine Aussage darüber getroffen werden kann, was

welchen Anteil am Effekt der Bedingung Inferenzprompts-via-Testleiter ausmacht. Aber negative Auswirkungen durch die Modalität(en) der Feedbackübermittlung in der Testleiterbedingung sind sehr wahrscheinlich auszuschließen. Welche Modalität der Feedbackdarstellung im Einzelfall (wann) genutzt wird, hängt vermutlich nicht nur von individuellen Vorlieben oder Fähigkeiten ab. Auch die Aufgabenkomplexität oder die Antwortsicherheit des Probanden können plausibler Weise auch Einfluss haben, so dass die Nutzung der einen oder der anderen Darstellungsvariante auch intraindividuell variieren mag.

Anhaltspunkte aus den Daten des Experiments

Die meisten Annahmen über die Wirkung des Testleiters als Feedbackgeber müssen letztlich Spekulation bleiben. Aus den gewonnenen Daten des Experiments sind jedoch zwei Aspekte herauszuheben, die zumindest dafür sprechen, dass die Probanden in der testleitervermittelten Feedbackintervention anders gearbeitet bzw. sich mehr angestrengt haben.

Die subjektive Einschätzung beim Instrument Anstrengungsthermometer ist es allerdings nicht, was zunächst kontraintuitiv erscheinen mag. Die Auswertung des Instruments zeigt nämlich, dass sich die Versuchsgruppen in der Einschätzung ihrer Anstrengung beim Bearbeiten des Tests in der Treatmentphase nicht unterschieden. Inwiefern die subjektiven, anonym abgegebenen Einschätzungen der eigenen Leistungsfähigkeit bzw. des Testverhaltens der Realität entsprachen, ist natürlich nicht festzumachen. Aber es spricht auch nichts dagegen, davon auszugehen, dass sich die Probanden im Allgemeinen und im Rahmen ihrer Möglichkeiten bemühten.

Stattdessen sprechen die Bearbeitungszeiten der Zweitantworten dafür, dass der Testleiter als Feedbackgeber tatsächlich (auch) die Anstrengungsbereitschaft (bei der Umsetzung der Rückmeldungen) erhöht hat. Die Testleiterbedingung hat für eine durchschnittliche Zweitantwort annähernd 39 Sekunden und damit ungefähr das Dreifache an der Zeit aufgewendet, die in den beiden anderen, erfolglosen Bedingungen Knowledge of Result und Inferenzprompt aufgebracht wurde.

Da die Bearbeitungsdauer für sich genommen keine inhaltliche Bedeutung hat, könnte die längere Bearbeitungszeit auch Symptom ungünstiger Verhaltensweisen/Prozesse beim Textverstehen sein, beispielsweise langsames Lesen oder ein ineffizienter

Strategieeinsatz. Andererseits treten die längeren Antwortzeiten eben in der Versuchsgruppe auf, die die Untersuchung am erfolgreichsten bestritt. Deshalb ist es naheliegend anzunehmen, dass durch die Intervention (meta-)kognitive Prozesse angeregt wurden, die zu einer (genaueren) Auseinandersetzung mit dem Material und damit zu einer längeren Bearbeitungsdauer in der Treatmentphase führten.
Weiterhin ist zu berücksichtigen, dass der Effekt der Bedingung Inferenzprompts-via-Testleiter eher nicht auf einen generellen Leistungsschub, ausgelöst etwa durch die Anwesenheit des Testleiters, zurückzuführen ist. Zwar bietet das unvollständige Design der Studie keinen Kontrast, der den reinen Effekt der Anwesenheit des Testleiters abbilden würde (vgl. Abschnitt 12.3): Aber die Analyse der Erstantworten, aufgesplittert in die einzelnen Units der Treatmentphase, belegte, dass sich die Überlegenheit der Testleiterbedingung erst später im Treatment in der Leistung niederschlug. Eine generelle Anhebung der Leistung, die nicht aus der eigentlichen Intervention resultierte, hätte sich von Anfang an zeigen müssen. Dieses Ergebnis wird dahingehend interpretiert, dass über das Setting mit dem Testleiter als Feedbackgeber (die) Rahmenbedingungen geschaffen werden, in denen die Prompts genutzt werden und deshalb wirken können.

Direkt vergleichbare Befunde sind aus der Feedbackliteratur nicht bekannt. Personen, Testleitern oder zum Beispiel Lehrer, werden durchaus häufiger als Feedbackgeber eingesetzt. Aber dabei wird die persönliche Feedbackgabe typischerweise nicht explizit als Notwendigkeit benannt, damit elaboriertes Feedback wirken kann. Stattdessen dient sie eher als passender Rahmen, in dem die Intervention mittels „natürlicher“ Interaktion stattfinden soll. Studien, die diese Variante der Feedbackpräsentation einsetzen, sind häufig auch als mehrmalige Interventionssitzungen oder eingebettet in ein Training konzeptualisiert. Der Untersuchungskontext erinnert dann an Tutoring-Lernsituationen.
Lernen in Interaktion mit (menschlichen) Tutoren gilt als ein sehr effektives Mittel, um Lernen und Leistung zu fördern. Der Vorteil des Tutoring hängt wesentlich mit dem Dialog (Dialogstruktur) und bestimmten Techniken der Interaktion bzw. Intervention durch den Tutor zusammen (Boyer et al., 2011). Aber mit dem Design/dem Setting der vorliegenden Untersuchung decken sich die Tutoring-Ansätze weniger. Über die testleitergebundene Präsentation der Rückmeldungen hinaus findet kein Dialog oder explizite Interaktion wie beispielsweise Zeigen relevanter Informationen im Text statt.
Am ehesten findet sich eine vergleichbare Testsituation bei Kluger und Adler (1993), obwohl deren Aussage bezüglich der personengebundenen Feedbackgabe kritisch

ausfallen und unter anderem deshalb im ersten Experiment die computervermittelte anstelle einer testleitergebundenen Feedbackgabe gewählt wurde. Die Arbeit von Kluger und Adler zeigt, dass die Testleistung durch die Präsentation von Feedback durch eine Person negativ beeinflusst wird (vgl. Abschnitt 3.5.1.3). Die Erklärungsansätze von Kluger und Adler, aber auch anderen (z.B. Ashford & Cummings, 1983; Comer, 2007) beziehen sich darauf, dass das Erhalten von Feedback durch eine unmittelbar anwesende, beobachtende Person als einschüchternd erlebt wird und deshalb abwehrende Prozesse auslöst und/oder aufgabenirrelevante Kognitionen befeuert, die sich nachteilig auf die Leistung auswirken. Bezogen auf die Leistung der Testleitergruppe in dieser Untersuchung wären diese Mechanismen integrierbar, wie im Zusammenhang mit den Ergebnissen zu den Erstantworten besprochen wurde.

Fazit zum „Testleitereffekt“

Die Überlegungen bezüglich der Quellen des „Testleitereffekts“ zusammengenommen ist davon auszugehen, dass durch den Testleiter als Feedbackgeber in erster Linie ein soziales Setting aufgebaut wird, das dem Probanden ein „Miteinander“ oder „Gegenüber“ schafft. Dies hat sehr wahrscheinlich wiederum soziale Prozesse, insbesondere sozialer Erwünschtheit ausgelöst, die unter anderem die Anstrengungsbereitschaft der Probanden in der Umsetzung der Rückmeldungen (Dauer der Auseinandersetzung mit Zweitversuchen) angehoben hat. Durch die Anwesenheit des Testleiters wird ein Ausweichen bzw. Vermeiden der Interventionen auch eher verhindert. Der Einfluss des Testleiters in der Wirkung des elaborierten Feedbacks wird also darin vermutet, dass er die Rahmenbedingung schafft, in denen die Wahrscheinlichkeit der Umsetzung der Rückmeldungen erhöht wird, wodurch die Inferenzprompts ihre Wirksamkeit für das Textverstehen entfalten können.

12.3 Überlegungen zum Nutzen der Inferenzprompts

Die wesentliche Erkenntnis des Experiments besteht darin, dass das elaborierte Feedback der Art Inferenzprompt sich – im Rahmen der testleitergebundenen Präsentation und über reine Präsenzeffekte des Testleiters hinaus – als förderlich für das Textverständnis erwiesen hat. Die Prompts vermitteln in Anpassung auf die jeweilige Aufgabenstellung

die Informationen, wie die für die Bewältigung der Aufgabenstellungen geforderten Inferenzen generiert werden können. Je nachdem welche Informationen (Agenten, Ereignisse) für eine Fragestellung bzw. Inferenz relevant sind, können die Hinweise auf textbasierte und/oder wissensbasierte Such- und Schlussfolgerungsprozesse ausgerichtet sein. Der Inhalt des Feedbacks setzt also konkret an den Prozessen an, die im Rahmen der psychologischen Theorien der Textverarbeitung (vgl. Abschnitte 2.1 und 2.2) als entscheidend für das Generieren von Inferenzen und damit als wesentliche Voraussetzung für den Aufbau einer kohärenten mentalen Repräsentation des zugrundeliegenden Sachverhalts angesehen werden.

Die Befunde lassen zunächst den Schluss zu, dass die spezifischen Informationen der Prompts dazu nützen, den Aspekt oder Ausschnitt der mentalen Repräsentation, der durch die vorliegende Aufgabe angesprochen ist, zu korrigieren bzw. überhaupt erst aufzubauen. Das belegen die signifikant höheren Leistungen in den Zweitversuchen, die das Gelingen der Korrektur einer Repräsentation reflektieren. Im Verlauf einer Unit ist das Wirken der Prompts auf dieser Ebene also sozusagen ein „puzzleartiges“ Vorgehen. Wenn eine Verknüpfung zwischen Informationen im Text und/oder dem repräsentierten Modell nicht (akkurat) gelingt, kann das Feedback genutzt werden, um sie herzustellen. Inwieweit sich diese „Reparaturarbeiten“ positiv auf die Konstruktion des gesamten Modells auswirken, also ob durch das punktuelle Eingreifen und Korrigieren als „Nebeneffekt“ auch andere Aspekte des Modells aktualisiert oder generiert werden, ist zu hinterfragen. Hierzu kann das vorliegende Experiment nur eine Aussage darüber machen, inwieweit es die weiteren Fragen einer Unit (derselbe Sachverhalt) betrifft und da zeigen sich die Auswirkungen eher von beschränktem Ausmaß. Denn im Vergleich zu den Kontrollbedingungen konnten die Inferenzprompts-via-Testleiter nur im kleineren Umfang und statistisch nicht bedeutsam mehr Aufgaben bereits im Erstversuch, die für die Transferierbarkeit der Intervention sprechen, beantworten. Darüber hinaus sind „Nebeneffekte“ der Korrektur einer Repräsentation infolge der Inferenzprompts aber nicht ausgeschlossen, weil einzelne (Multiple-Choice-) Aufgaben keine Aussagen über die konstruierten Modelle als Ganzes erlauben.

Dass die Transferleistung aus den Inferenzprompts weniger ad-hoc funktionierte als die Korrektur derselben Aufgabe, wird der Komplexität der Anforderung zugeschrieben, die

zudem durch den Kontext der Untersuchungssituation (d.h. einmalige Intervention, hoher Arbeitsaufwand) eher noch erschwert wird.

Neben dem ressourcenfordernden Charakter der Inferenzprompts kann auch die inhaltliche Spezifität kritisch für die Transferleistung sein: die Spezifität der Intervention ist einerseits ein Vorteil über den die Rückmeldungen verfügen. Spezifität von Feedback gilt als eines der wesentlichen erfolgversprechenden Merkmale (Shute, 2008). Es bedeutet aber auch fehlende Generalität. Durch die spezifische Ausrichtung von Rückmeldungen auf die Aufgabenstellungen müssen, wenn es keine weiteren, transferfördernden Hinweise gibt, generellere Prinzipien durch den Leser erschlossen werden. Das gilt wohl umso mehr im Bereich des Textverstehens. Dass Transferleistungen oft nur schwer zu erreichen sind, zeigt nicht nur das vorliegende Experiment (vgl. Bangert-Drowns et al., 1991). Würden aber vice versa nur generelle, allgemeine Prinzipien für erfolgreiches Textverstehen vermittelt, fehlten wiederum die konkreten Ausrichtungen für die jeweiligen Anforderungen.

Bereits angesprochen wurde die Generalisierbarkeit der Inferenzprompts. Aus der Perspektive der testleitergebundenen Feedbackpräsentation wird die Generalisierbarkeit der Inferenzprompts darin gesehen, dass sie in Lehr-Lernsituationen, in denen das Feedback durch eine anwesende Person gegeben und seine Umsetzung beobachtet wird, eingebaut werden können. Das betrifft also vor allem die natürlichen Lernsettings im schulischen Unterricht oder ähnlichen Kontexten. Die Wirksamkeit der Prompts könnte dann vermutlich dadurch erhöht werden, dass zwischen beispielsweise Schüler und Lehrer ein Dialog über oder auf der Grundlage der Prompts geführt werden kann. Das Verständnis und das Wissen über strategische Textverarbeitung könnten auf diesem Weg ausgebaut und gefestigt werden. Auch weiterführende Hinweise, etwa zur Generalisierung der in den Rückmeldungen konkret beschriebenen kognitiven Prozesse der Textverarbeitung könnten gut an die Arbeit mit den Rückmeldungen angeknüpft werden.

Nun wurden im vorliegenden im Unterschied zum vorherigen Experiment die Schulformen Hauptschule und Gymnasium bei der Stichprobengewinnung nicht berücksichtigt. Insofern stellt sich die Frage, inwieweit der positive Effekt der Inferenzprompts-via-Testleiter auch für Schüler dieser beiden Schulformen zu erwarten ist: in der Feedbackliteratur finden sich vereinzelt Hinweise, dass Feedbackarten, besonders einfache versus elaborierte Arten, einen differenziellen Effekt in Abhängigkeit

vom Leistungslevel der Lerner haben könn(t)en. Dabei ist natürlich zu berücksichtigen, dass die Schulform allein kein ausreichendes Kriterium für die Leistungsfähigkeit der Probanden darstellt. Außerdem liegen für den Bereich des Textverstehens oder ähnlicher Anforderungsbereiche keine entsprechenden Hinweise vor. Deshalb ist in der Ableitung des Untersuchungskonzeptes für das erste Experiment keine entsprechende Forschungsfrage formuliert worden. Auch die Datenlage des ersten Experiments suggeriert hier keine differenziellen Effekte. Insofern ist davon auszugehen, dass auch Schüler der Hauptschule und der Gymnasien von den Inferenzprompts profitieren können.

Daneben stellt sich auch die Frage, inwiefern die Wirkung der Inferenzprompts auch auf Schüler höherer oder kleinerer Klassenstufen generalisierbar ist. Verständnisschwierigkeiten, die aus defizitären Informationsverarbeitungsprozessen auf der Textebene resultieren, sind für verschiedene Altersgruppen von der Grundschule bis ins Erwachsenenalter nachgewiesen (vgl. Allington & McGill-Franzen, 2009). Die grundlegenden Fähigkeiten zum verstehenden Lesen von Texten sind schon beherrschen Leser im Allgemeinen schon vor der sechsten Klassenstufe (vgl. Abschnitt 2.3). Die Anforderungen dabei ändern sich für verschiedene Altersgruppen eher durch die Schwierigkeit der Texte, insbesondere auch durch ihre Inhalte und dem Vor- und Weltwissen, das sie voraussetzen.
Insofern wird davon ausgegangen, dass das Prinzip der Inferenzprompts durchaus über die Gruppe der Sechstklässler hinaus generalisierbar ist; Leseanfänger ausgenommen. Wichtiger für das Gelingen der Intervention erscheint dann die Angemessenheit der eingesetzten Texte und Aufgabenstellungen, an die die Prompts dann angepasst sein müssen.

12.4 Diskussion der Untersuchungsmethodik

Die in diesem Experiment eingesetzten Texte und Lesekompetenzitems basieren auf den Materialien der ersten Untersuchung und sind daher (immer noch) Eigenkonstruktionen. Die Nutzung eigenkonstruierter Items im Experiment ist an anderen Stellen bereits diskutiert und das Vorgehen mit Blick auf das Anliegen der vorliegenden Arbeit als vertretbar eingestuft worden (vgl. Abschnitt 10.8.2 für Methodik dieses Experiments und

Abschnitt 7.3 für Diskussion zu Experiment 1). Dem ist hinzuzufügen, dass die Voraussetzungen bezüglich der Itemkennwerte im vorliegenden Experiment besser gewesen sind als im ersten Experiment – da eben durch das Administrieren der Items im ersten Experiment bereits eine erste Überprüfung der Items vorlag.
Das Material ist auf der Grundlage der Kennwerte des ersten Experiments für diese Untersuchung teilweise angepasst worden. Die Entscheidung für die Abänderungen am Material ändert nichts an der Gültigkeit der Aussagen zum ersten Experiment. Gleichzeitig resultiert daraus aber, dass eine unmittelbare Vergleichbarkeit der Leistungen zwischen spezifischen Units der einen und der anderen Untersuchung nicht geboten ist.
Darüber hinaus belegt die Analyse der Itemkennwerte anhand der Daten der Kontrollgruppe, dass die Lesekompetenzitems im Allgemeinen wiederum eine mittlere Schwierigkeit ($p = .51$) aufweisen. Damit ist, wie zuvor auch, die Grundlage für potentielle Leistungsentwicklungen geschaffen. Das Material ist im Allgemeinen weder zu schwierig, wodurch die Nutzbarkeit von Feedback für die Lerner außer Reichweite gebracht würde, noch ist es zu leicht, wodurch Feedbackinterventionen kaum oder gar nicht zum Einsatz gekommen wären.

Ein diskussionswürdiger Aspekt im Umgang mit den Daten ist die Bereinigung der Auswertungsstichprobe durch den Ausschluss der so genannten „Durchklicker“, also jener Fälle, die anhand festgesetzter Kriterien zu viele Items zu schnell beantwortet haben. Die zugrunde gelegten Kriterien stimmen mit dem Vorgehen im ersten Experiment überein. Dort sind die Notwendigkeit der Bereinigung sowie die Wahl der Kriterien und mögliche Auswirkungen auf die Analysen bereits dargelegt (vgl. Abschnitt 10.8.1 für Methodik dieses Experiments und Abschnitt 7.3 für Diskussion zu Experiment 1). Dem ist aus der Perspektive des vorliegenden Experiments nichts hinzuzufügen und es wird bezüglich dieser Aspekte auf die angegebenen Abschnitte verwiesen.

Als letzter Aspekt der Untersuchungsmethodik wird noch auf die Repräsentativität der Stichprobe eingegangen. Diese ist zum einen wiederum dadurch eingeschränkt, dass die Teilnahme an der Untersuchung zunächst für die angefragten Schulen und dann für jeden Schüler bzw. seine Erziehungsberechtigten freiwillig war. Zum anderen rekrutiert sich die Stichprobe nur aus Realschulen. Die Beschränkung auf eine Schulform war dem Umstand geschuldet, dass dieses Experiment nur in reduzierterem Umfang durchgeführt werden

konnte. Die Wahl der Schulform Realschule begründet sich dadurch, dass hier, die Schulform als grober Indikator für die allgemeine Leistungsfähigkeit der Schüler, in Relation zu Hauptschulen und Gymnasien eher ein mittlerer Leistungsbereich zu erwarten ist, der die beste Passung zu den vorhandenen Materialien versprach. Wie für die Lesekompetenzitems bereits erläutert, wird davon ausgegangen, dass sie im Rahmen von Feedbackinterventionen keine extreme Schwierigkeit aufweisen sollten. Durch die Beschränkung auf eine Schulform ist die direkte Vergleichbarkeit der empirisch gefundenen Effekte beider Experimente allerdings nicht gegeben.

12.5 Gesamtfazit des Experiments

Den Testleiter als Feedbackgeber einzusetzen hat sich im Zusammenhang mit dem elaborierten Feedback der Art Inferenzprompt als effektiv für das Textverständnis/die Lesekompetenz erwiesen. Der Effekt der Intervention geht nicht (nur) auf die Präsenz des Testleiters, etwa im Sinne eines generellen Leistungsanstiegs, zurück. Sondern es ist davon auszugehen, dass die Inferenzprompts den wesentlichen Anteil an der Leistungssteigerung haben. Inhaltlich zeichnen sich die Prompts dadurch aus, dass sie die zur Bewältigung einer Anforderung notwendigen text- und/oder wissensbasierten Such- und Schlussfolgerungsprozesse benennen. Der Effekt des Testleiters kann darauf zurückgeführt werden, dass er die Rahmenbedingung schafft, in denen die Wahrscheinlichkeit der Umsetzung der Rückmeldungen erhöht wird, wodurch die Inferenzprompts ihre Wirksamkeit für das Textverstehen entfalten können. Dass eine Intervention auf dieser Ebene notwendig sein könnte, wurde aus den Erkenntnissen des ersten Experiments abgeleitet, und nun zum einen durch den Effekt der Testleiterbedingung und zum anderen aufgrund der wiederum ausbleibenden Effekte für die rein computerbasierte Darbietung der Inferenzprompts bestätigt.

13 Abschließende Diskussion beider Experimente und Ausblick

Die übergreifende Fragestellung beider Experimente war auf die Wirksamkeit von Feedbackinterventionen auf das Textverstehen/die Lesekompetenz gerichtet, unter den spezifischen Rahmenbedingungen, die sich auch für den über diese Arbeit hinaus geplanten dynamischen Kurzzeitlerntest ergeben. Die Inhalte beider Experimente lassen sich wie folgt kurz zusammenzufassen. Das erste Experiment fokussierte drei elaborierte Feedbackarten verschiedenen Inhalts: die Fehlererklärung, der metakognitive Prompt und die Inferenzprompts. Das übergreifende Ergebnis bestand darin, dass keine der Feedbackinterventionen eine Leistungssteigerung nach sich zog, weder in der Treatmentphase noch in den beiden Posttests. Die Diskussion möglicher Gründe für die Ineffektivität der Feedbackinterventionen mündete in zwei konkurrierenden Annahmen: entweder die genutzten elaborierten Feedbackinhalte sind für das Textverstehen unbrauchbar oder sie wurden von den Probanden aufgrund unzureichender Anstrengungsbereitschaft (Motivation) nicht verarbeitet.

Unter Berücksichtigung des Gesamtkontextes wurde der letzeren Hypothese mit dem zweiten Experiment nachgegangen. Der Ansatz bestand dabei darin, über die Feedbackgabe durch den Testleiter in einer Eins-zu-Eins-Sitzung mit dem Probanden ein Setting zu schaffen, das sich so in der Durchführung der Kognitiven Interviews bewährt hatte, und über das eine ausreichende Anstrengungsbereitschaft gewährleistet werden sollte. Die Intervention mit testleitergebundener Feedbackgabe wurde mit dem elaborierten Feedback Inferenzprompts durchgeführt, da ihnen aus theoretischer Perspektive das höchste Potential für das Textverstehen zugesprochen wird. Das zentrale Ergebnis des zweiten Experiments bestand darin, dass die Inferenzprompts, wenn sie über den Testleiter gegeben wurden, gegenüber den Kontrollbedingungen eine bedeutsame Leistungssteigerung sowohl in den Zweitantworten als auch im Posttest bewirkten. Dabei ist davon auszugehen, dass die Leistungssteigerung nicht auf eine generelle Anhebung der Performanz (nur) durch die Präsenz des Testleiters zurückgeht, sondern im Wesentlichen auf die Inferenzprompts.

Übergreifende Betrachtung der Befunde beider Experimente

Das zweite Experiment hat neben der neuen testleitergebundenen Feedbackintervention auch zwei Interventionen des ersten Experiments wiederholt, und zwar die ausschließlich computerbasierte Präsentation von Inferenzprompts einerseits und von Knowledge of Result andererseits. In beiden Experimenten stellten sich diese Bedingungen als wirkungslos zur Steigerung der Leistung heraus. In dieser Hinsicht repliziert das zweite Experiment das erste und untermauert den Befund zur Ineffektivität der computerbasierten Gabe von elaboriertem Feedback im Rahmen des vorliegenden Testsetting.

Bezüglich der Bedingung Knowledge of Result wird davon ausgegangen, dass hier die Feedbackart selbst ineffektiv ist, die ausschließlich computerbasierte Übermittlung also nicht ausschlaggebend für die Ineffektivität der Bedingung war. Für das Feedback Knowledge of Result ist auch unter Einbezug der bestehenden Feedbackliteratur (vgl. Abschnitt 3.5.1.1) davon auszugehen, dass es im Allgemeinen nicht zur Verbesserung des aktuellen Textverständnisses und/oder der Lesekompetenz genutzt werden kann. Sein Inhalt bietet keine Hilfestellungen. Das Rückmelden, dass ein Fehler vorliegt, stimuliert keine weiterführenden, verständnis- bzw. leistungsförderlichen Aktivitäten der Textverarbeitung. Für bestimmte Fähigkeitsbereiche (vgl. Abschnitt 3.3) oder bei hohem Vorwissen (Hanna, 1976; vgl. auch Abschnitt 3.5.2) kann diese einfachste Art des Feedbacks unter Umständen zwar gewinnbringend genutzt werden, indem es letztlich zu mehr Anstrengung und/oder dem Einsatz effektiverer Strategien führt (Song & Keller, 2001; vgl. Abschnitt 3.3). Aber um Verständnisschwierigkeiten zu überwinden, ist es im Allgemeinen nicht brauchbar, wie auch die in Abschnitt 3.5.1.1 referierten Feedbackstudien im Bereich des Textverstehens belegen. Es bietet dafür zu wenig Information und keine ausreichende Form der Hilfestellung. Um Verständnislücken zu schließen oder falsch konstruierte Relationen im Situationsmodell zu korrigieren, werden konkretere Hilfen benötigt. Davon auszuschließen sind selbstverständlich Flüchtigkeitsfehler und falsche Antworten, die eine (letztlich die falsche) Alternative von mindestens zwei war, zwischen denen der Proband geschwankt hatte. Hier ist der Hinweis, dass die abgegebene Antwort falsch ist, sicherlich ausreichend und die richtige Lösung kann dann umgehend genannt werden.

Die konkreten Hilfen sind dagegen in den elaborierten Feedbacks enthalten, bei denen, wie die beiden Experimente dieser Arbeit zeigen, die Art ihrer Übermittlung bzw. Einbettung in den Untersuchungskontext Bedeutung zukommt. Im zweiten Experiment zeigte das der direkte Vergleich der beiden Bedingungen mit Inferenzprompts, und zwar als rein computerbasierte Intervention einerseits und über den Testleiter vermittelt andererseits. Die Inferenzprompts-via-Testleiter erwiesen sich nicht nur als die wirksamste Intervention für das Textverstehen. Sondern die ausschließlich computerbasierte Intervention mit Inferenzprompts brachte, gemessen an den Leistungen der Kontrollbedingungen (kein Feedback, Knowledge of Result), keinen Effekt hervor und zeigte sich im Vergleich zur Testleiterbedingung auch signifikant schlechter. Das bedeutet, die computerbasierte Intervention mit Inferenzprompts konnte von den Probanden nicht für die Beantwortung und Korrektur der Verständnisfragen genutzt werden – sie erzielten eine Leistung auf dem Niveau der Testbedingung ohne Feedback und der Intervention mit dem ebenfalls nicht wirksamen Knowledge of Result. Von diesen drei Bedingungen schnitt die computerbasierte Gabe von Inferenzprompts eher noch am schlechtesten ab, da allein ihr Kontrast zur Testleiterbedingung statistische Bedeutsamkeit in den Erstantworten erreichte. Die Ineffektivität der rein computerbasierten Darbietung der Inferenzprompts zeigte sich auch im ersten Experiment, in dem ebenfalls die beiden anderen, wirkungslos gebliebenen elaborierten Feedbacks, Fehlererklärung und metakognitiver Prompt, als ausschließlich über den Computer vermittelte Feedbackinterventionen umgesetzt wurden.

Der Befund, dass die Inferenzprompts-via-Testleiter zu einer Leistungssteigerung führten, die computerbasiert übermittelte Variante aber nicht, gibt ebenso eine Antwort auf die in der Diskussion des ersten Experimentes aufgestellten Hypothesen hinsichtlich der Ursachen der Ineffektivität der computerbasierten Feedbackbedingungen. Als mögliche Erklärung der Ineffektivität wurde vermutet, dass entweder die Inhalte der Feedbacks unbrauchbar für das Textverstehen waren oder die Feedbacks aufgrund unzureichender Anstrengung/Motivation nicht verarbeitet wurden. Das zweite Experiment spricht klar für die Hypothese der unzureichenden Motivation zur Umsetzung der Rückmeldungen in der computerbasierten Intervention mit Inferenzprompts. Denn der Testleiter in der Bedingung Inferenzprompts-via-Testleiter hat keine weiteren Hilfestellungen gegeben und die Inferenzprompts waren dieselben wie in der rein computerbasierten Bedingung.

Die möglichen Gründe bzw. Umstände, die dazu führten, dass die elaborierten Rückmeldungen in den computerbasierten Interventionen nicht umgesetzt wurden, sollen an dieser Stelle auch in Hinblick auf die noch zu erläuternden Implikationen für den Dynamischen Kurzzeitlerntest noch einmal zusammengefasst werden. Auf der einen Seite bestand die Untersuchung in der Umsetzung eines kognitiv anspruchsvollen Programms, indem mehrere Texte mit einer Reihe von Aufgaben bewältigt werden mussten, was allein schon einen relativ hohen Leseaufwand bedeutete. Zudem sind die Aufgaben, die in erster Linie das Verständnis auf Textebene prüfen, eher nicht leicht zu beantworten. Es handelt sich um einen Fähigkeitsbereich, der vielen Lesern eher nicht leicht fällt. Das zeigen auch in beiden Experimenten die relativ hohen durchschnittlichen Fehlerquoten von 54 % bis 58 % in den Erstantworten der Treatmentphase.

Aber genau diese Fehler werden für Feedbackinterventionen benötigt, da die Intervention an den Fehlern ansetzt. Es sollte umso eher von Rückmeldungen gelernt werden können, wenn sie öfter präsentiert werden, das hängt aber wiederum vom Auftreten von Fehlern ab. Viele Fehler zu machen, könnte auf Seiten des Lerners aber zu Frustration oder Überforderung geführt haben, was in der Folge das Wirken der Intervention in der Kürze sehr schwierig macht.

Auf der anderen Seite ist zu berücksichtigen, dass die Teilnahme bzw. das Abschneiden beim Experiment an keine Konsequenzen (Noten oder ähnliches) für die Schüler gebunden war. Die Inhalte der Untersuchung sind in gewisser Weise auch dekontextualisiert, das heißt, die Sinnhaftigkeit des Test-Lern-Test-Prozedere mag den Schüler nicht oder nur eingeschränkt zugänglich sein. Der Nutzen der Prozedur war ihnen vermutlich nicht gut vermittelt, aber im Rahmen einer Testung von extern und losgelöst vom Unterrichtsgeschehen auch nur begrenzt möglich.

Es kann also in Bezug auf die Erwartungs-Wert-Modelle der Motivationspsychologie, die in diesem Zusammenhang in Abschnitt 7.2.2 zur Erklärung herangezogen wurden, davon ausgegangen werden, dass die fehlende bzw. unzureichende Motivation zur Umsetzung der Rückmeldungen sowohl auf die Wertkomponente als auch die Erwartungskomponente zurückgehen kann. Hinzu kommt, dass die Schüler bei der computerbasierten Darbietung zwar in einer Gruppentestungssituation arbeiteten, in der ebenfalls ein Testleiter anwesend war, aber letztlich bearbeitete jeder Proband das Programm am Computer für sich selbst, anonym. Dieser Umstand kann, gerade auch in Abgrenzung zur Testleiterbedingung, ein Vermeiden der Feedbackverarbeitung erleichtert haben.

In der Testleiterbedingung waren die Anforderungen dieselben und entsprechend auch die Erwartungen, die die Schüler hinsichtlich der Änderbarkeit und Umsetzbarkeit der Rückmeldungen hatten. Auch an den kritischen Punkten zum erkennbaren Nutzen der Feedbackinterventionen änderte die Testleiterbedingung wenig. Aber das Setting der Testung war stärker hin zu einem sozial ausgerichteten Setting ausgerichtet, auch wenn kein Dialog und echter Austausch stattfand. Aber der Proband erhielt ein Gegenüber, der ihm die Rückmeldungen gab. Dieser Umstand und vermutlich vor allem soziale Prozesse wie soziale Erwünschtheit verhinderten oder erschwerten zumindest das Vermeiden der Umsetzung der Rückmeldungen. Durch die Schaffung dieser Rahmenbedingungen konnten, wie in der Diskussion des zweiten Experiments geschlussfolgert wurde (vgl. Abschnitt 12), die Inferenzprompts auch ihren Nutzen entfalten. Sie führten zu einer Verbesserung des Textverständnisses, sowohl in der Treatmentphase als auch im Posttest. Das heißt, die Probanden dieser Bedingung konnten anhand der Inferenzprompts mehr falsche Antworten korrigieren und die Informationen der Rückmeldungen auch auf nachfolgende Aufgaben transferieren, so dass mehr Aufgaben schon im ersten Versuch bzw. im Posttest richtig beantwortet wurden.

Die Höhe der Effekte der Inferenzprompts-via-Testleiter belief sich immer auf mittelgroße Effekte (zwischen Cohens $d = 0.49$ und $d = 0.67$). In der Arbeit von Hattie (2009), in der Metaanalysen zu verschiedensten Einflüssen auf die schulische Leistung aggregiert sind, wird der durchschnittliche Effekt von Feedback mit ES = 0.73 angegeben und rangiert damit auf dem zehnten Platz der wirksamsten Mittel zur Steigerung der Leistung. Hattie stuft in seiner Arbeit Effekte ab $d = 0.40$ als wünschenswerte Effekte ein. Vor diesem Hintergrund und unter Berücksichtigung der Komplexität der Anforderungen und den spezifischen, eingeschränkten Rahmenbedingungen der vorliegenden Untersuchung wird die erzielte Leistungssteigerung der Inferenzprompts-via-Testleiter als sehr positiv bewertet.

Die Gabe der Inferenzprompts über den Testleiter stellte sich also als erfolgreiche Intervention dar. Aber die Ergebnisse diese Bedingung schließen nicht aus, dass die Feedbackgabe über den Testleiter auch nachteilige Auswirkungen auf die Probanden hatte und in der Folge möglicherweise die Leistungsfähigkeit unter den Inferenzprompts (in der Treatmentphase) abgeschwächt hat. Die Risiken, die mit einer personengebundenen Feedbackgabe einhergehen, insbesondere wenn der Feedbackgeber der Testleiter ist, wurden im theoretischen Teil der Arbeit (vgl. Abschnitt 3.5.1.3) dargelegt. Eine Abschwächung der Leistung tritt dann ein, wenn bei den Probanden aufgabenirrelevante

Kognitionen verstärkt werden. Dafür spricht, dass der Leistungsvorsprung der Bedingung Inferenzprompts-via-Testleiter in der Treatmentphase weniger deutlich ausfällt als im Posttest, in dem der Testleiter zwar noch anwesend, die Feedbackprozedur aber beendet war. Das schmälert nicht die Befunde des Experiments, ist aber bei den Überlegungen zur Umsetzung der Feedbackinterventionen im Rahmen weiterer Untersuchungen, im Speziellen für den Dynamischen Kurzzeitlerntest zu berücksichtigen.

Fazit: Beide Experimente zusammengenommen demonstrieren die Anforderungen, die der Einsatz elaborierter Feedbacks in einer Kurzzeitintervention zur Unterstützung des Textverstehens bzw. zur Überwindung von Verständnisschwierigkeiten auf der Textebene mit sich bringt. Die Befunde betonen dabei den Faktor der Feedbackrezeption. Darüber hinaus belegt die Arbeit, dass Inferenzprompts, also kognitive Hinweise zum Herstellen einer geforderten Inferenz, auch unter den spezifischen Rahmenbedingungen dieser Arbeit nützlich und hilfereich sind, um a) eine erfragte, aber zunächst falsch hergestellte Inferenz erfolgreich zu korrigieren, b) die über die Rückmeldungen vermittelten Informationen im Verlauf des Treatments bereits auch auf folgende Antworten im Erstversuch anzuwenden, was sich c) aber vor allem im Posttest manifestierte.

Generalisierbarkeit der Befunde

Eine erste Frage der Generalisierbarkeit der Befunde bezieht sich auf die beiden computerbasierten Feedbackinterventionen Fehlererklärung und metakognitiver Prompt des ersten Experiments und ob oder inwieweit für sie positive Effekte auf die Leistung zu erwarten sind, wenn sie über den Testleiter vermittelt werden würden. Wie im Abschnitt 3.5.3 unter Nutzung der Feedbackbefunde und der Theorien des Textverstehens hergeleitet, wird davon ausgegangen, dass auch die Feedbacks Fehlererklärung und metakognitiver Prompt einen Beitrag zur Förderung des Textverständnisses leisten können. Die Voraussetzungen ihres Wirkens werden im Vergleich zu den Inferenzprompts hingegen als höher eingestuft, da sie stärker den selbstregulierten Einsatz von angemessenen Such- und Regulationsprozessen und das Vorhandensein des entsprechenden Wissens erfordern. Durch die Kombination mit dem testleitergebundenen Präsentationsmodus sind Effekte auf die Leistung zu erwarten, in Anbetracht der höheren

kognitiven Voraussetzungen sollten sie allerdings kleiner ausfallen als die gefundenen Effekte der Inferenzprompts-via-Testleiter.

Eine zweite Frage zur Generalisierbarkeit der Befunde bezieht sich auf die Anwendungen der Inferenzprompts in Lernsituationen jenseits des speziellen Test-Lern-Settings dieser Arbeit. Entsprechende Überlegungen wurden bereits ausführlich in Abschnitt 12.3 dargelegt und werden an dieser Stelle kurz wiedergegeben: Es ist davon auszugehen, dass sich die Inferenzprompts, wenn sie schon in dem stark eingeschränkten Setting der Untersuchung dieser Arbeit wirken, dann auch in natürlicheren (schulischen) Settings wirkungsvoll erweisen. Im Unterricht etwa können die Prompts für Verständnisfragen eingesetzt werden, um den Schülern bei der Lösung der Aufgaben und letztlich beim Aufbauen eines tieferen Textverständnisses Hilfestellung zu geben. Als Ergänzung bieten sich auch weiterführende Erklärungen an, etwa welche Textstellen relevant sind, oder warum eine Antwort falsch ist. Die Ergänzung der Inferenzprompts durch andere Hilfestellungen entspricht auch dem Vorgehen bei der erfolgreichen Intervention von Winne und Kollegen (1993; vgl. Abschnitt 3.5.1.1).
Auch über die Altersgruppe der 11- bis 12-Jährigen (Sechstklässler), die in dieser Arbeit untersucht wurden, hinaus, spricht nichts gegen den Nutzen der Inferenzprompts. Vor allem auf ältere Leser ist der Effekt generalisierbar. Für jüngere Leser ist eine Wirkung ebenfalls nicht ausgeschlossen, bei Winne und Kollegen (1993) wurde beispielsweise bei Dritt- bis Fünftklässlern erfolgreich interveniert, allerdings bestand die Feedbackintervention auch aus zusätzlichen Erklärungen und die Intervention verteilte sich auf mehrere Sitzungen. Ebenso wird erwartet, dass sich die Effekte der Inferenzprompts-via-Testleiter auch bei Schülern der Hauptschulen und Gymnasien wiederholen ließen (vgl. Abschnitt 12.3).

Eine dritte Frage zur Generalisierbarkeit der Befunde bezieht sich auf den „Testleitereffekt“ und wie dieser, insbesondere auch in Hinblick auf die Entwicklung des Dynamischen Kurzzeitlerntests, in Untersuchungssituationen mit einer Gruppensitzung installiert werden kann. Zwar erwies sich die testleitergebundene Gabe von Inferenzprompts als effektive Maßnahme, aber daneben ist zu klären, wie es gelingen kann, den Testleitereffekt auf einem anderen Weg in Testsituationen zu integrieren. Denn zum einen ist mit der Feedbackgabe über den Testleiter ein vergleichsweise hoher Aufwand in der Untersuchungsdurchführung verbunden und zum anderen sind mögliche,

den Effekt auf die Leistung mildernde Auswirkungen nicht ausgeschlossen. Außerdem spricht es nicht für eine Feedbackintervention, wenn sie ausschließlich vermittelt über eine unmittelbar anwesende Person funktioniert.

Ausgehend von den Überlegungen zu dem Wirken der testleitergebundenen Feedbackgabe kann der „Testleitereffekt“ in ähnlicher Weise auch durch das Anbinden der Ergebnisse im (feedbackgestützten) Test an relevante Konsequenzen geschaffen werden. Wenn das Abschneiden im Test beispielsweise an Noten gekoppelt wird, ist davon auszugehen, dass die Bedeutung der Tests und vor allem der Umsetzung der Rückmeldungen erhöht wird. In der Folge kann erwartet werden, dass eine hinreichende Anstrengungsmotivation aufgebracht wird.

Ein anderer Ansatz ist die Abschwächung der Dekontextualisierung der Inhalte der Untersuchung. Über die Einbindung von schulrelevanten Texten und Verständnisabfragen (Aufgaben) oder den Interessen der Schüler entsprechendem Material könnte der Nutzen für die Probanden erhöht und die Bereitschaft zur Umsetzung der Rückmeldungen gesteigert werden.

Ein weiterer Ansatz liegt in der Veränderung des Computerprogramms. Die Testleiterbedingung zeichnet sich nicht nur durch die Schaffung eines gewissen sozialen Drucks aus (Testbearbeitung im Sinne der sozialen Erwünschtheit), sondern durch die testleitergebundene Feedbackgabe ändert sich für den Lerner auch die Situation der Feedbackübermittlung. Er erhält ein Gegenüber, der ihm die Rückmeldungen gibt und der der Interventionssituation auch eine Bedeutung gibt. Übertragen auf das Computerprogramm könnte sich beispielsweise die Implementierung einer animierten Figur, die die Rückmeldungen verbal (über Audiofiles) gibt, bereits positiv auswirken. Der Einsatz animierter Figuren (*paedagogical agents*) ist in computerbasierten Lernumgebungen weit verbreitet und auf mehr oder weniger Interaktivität mit dem Lerner ausgelegt (Graesser, Jeon & Dufty, 2008; Graesser & McNamara, 2010). Für das Untersuchungssetting dieser Arbeit könnte der Rückgriff auf eine animierte Figur als „Feedbackgeber“ das Gegenüber für die Lerner schaffen, das sie in der Bearbeitung der Rückmeldungen eine Bedeutung erkennen lässt bzw. ihnen einen Anreiz dazu liefert.

Kritik der Untersuchungsmethodik

Die beiden Experimente dieser Arbeit zeichnen sich im Vergleich zu den bestehenden Feedbackstudien im Bereich des Textverstehens dadurch aus, dass zur Erfassung des Textverständnisses nicht nur einzelne Textpassagen, sondern verhältnismäßig längere Texte genutzt wurden. Es wurden auch ausschließlich Aufgabenstellungen eingesetzt, die das Verständnis des Gelesenen (der Bedeutungsrepräsentation) abprüfen. Außerdem wurden verhältnismäßig viele Aufgaben zur Erfassung der Leistung vorgegeben.

Im ersten Experiment wurden drei elaborierte Feedbackarten untersucht, die vor dem Hintergrund der Befunde der Feedbackliteratur auch unter den spezifischen Rahmenbedingungen dieser Arbeit als potentiell wirksam für das Textverstehen eingeschätzt wurden. Das zweite Experiment griff dann zunächst nur die Inferenzprompts wieder auf, für die sich positive Effekte in der Testleiterbedingung nachweisen ließen. Darüber hinaus wird es aber auch von Bedeutung sein, die Effektivität weiterer elaborierter Feedbackarten, speziell der Fehlererklärung und des metakognitiven Prompts, unter den wirksamen Rahmenbedingungen (Testleiterbedingung oder Äquivalent) zu untersuchen. Die Effekte elaborierten Feedbacks für das Textverstehen würden weiter abgeklärt werden, indem die Wirksamkeit der Inferenzprompts auch mit anderen elaborierten Feedbackarten verglichen werden würde. In diesem Punkt konnte die Arbeit die Absicht des ersten Experiments nicht vollständig im zweiten Experiment realisieren.

Des Weiteren ist an der Untersuchungsmethodik hervorzuheben, dass zur Bewertung der Feedbackwirksamkeit verschiedene Indikatoren herangezogen wurden. Es wurde nicht nur die Transferleistung anhand eines Posttest erfasst, sondern auch die Leistung in der Treatmentphase, aufgeschlüsselt nach Erst- und Zweitantworten. Die Nutzung der Erst- und vor allem der Zweitantworten ist in erster Linie der Anpassung der Testbedingungen an den Dynamischen Kurzzeitlerntest (vgl. Abschnitt 3.5.3) geschuldet. Aber auch darüber hinaus ist die Berücksichtigung dieser Indikatoren aus der Perspektive der Feedbackliteratur interessant. Vor allem die Leistungen untermittelbar nach der Feedbackgabe wurden bisher (weitestgehend) nicht herangezogen, obwohl es sich dabei um einen wichtigen Indikator handelt (Kulhavy & Stock, 1989; vgl. Abschnitt 3.4).
Nichtsdestotrotz könnte in dieser Art des Wechsels von Intervention und Erfassung der Leistung auch die Gefahr der Überforderung der Probanden liegen. Insbesondere schwächere Leser, die eher viele Fehler machen, werden häufig unterbrochen und ihnen

werden wiederholt ihre Defizite vor Augen geführt. Diese Bearbeitungsbedingungen sind sicher ungewohnt, möglicherweise auch belastend für den Leser.

Bedeutung der Experimente für den geplanten Dynamischen Test

Die zentrale Erkenntnis dieser Arbeit für den geplanten Dynamischen Test in Form eines Kurzzeitlerntests besteht darin, dass Inferenzprompt auch unter den spezifischen Bedingungen des Tests eine effektive Maßnahme zur Unterstützung des Textverstehens/der Lesekompetenz darstellen. Die Effektivität zeigte sich vor allem in der Korrekturleistung falscher Erstantworten und in der Transferleistung im unmittelbar anschließenden Posttest. Der Effekt auf die Erstantworten der Treatmentphase fiel insofern geringer aus, dass er sich nicht gegenüber einer der Kontrollbedingungen (kein Feedback oder Knowledge of Result) zeigte, sondern gegenüber der rein computerbasierten Gabe von Inferenzprompts. Letzterer Aspekt weist auf die Herausforderungen der Feedbackinterventionen auf der Grundlage des Textverstehens bzw. von Verständnisschwierigkeiten auf Textebene hin. Zudem wurde die Wirksamkeit der Inferenzprompts bisher nur in der testleitergebundenen Übermittlung nachgewiesen.

Die Erkenntnisse des zweiten Experiments sind insofern unmittelbar auf die Entwicklung des Tests anwendbar, wenn dieser als Einzelsetting zu konzipieren wäre. Allerdings muss das Anliegen der Entwicklung des Dynamischen Tests in einer zuverlässig funktionierenden Gruppentestung bestehen. Schon die Praktikabilität weiterer Untersuchungen zur Entwicklung des Tests (z.B. Untersuchungen der Validität) erfordern einen ökonomisch durchzuführenden Gruppentest. Der Aufwand der testleitergebundenen Feedbackgabe in Einzelsettings ist dafür zu hoch. Hierfür wird insbesondere die Frage nach der Übertragbarkeit des Testleitereffekts in Testsettings ohne Einzelsitzung und ohne die Feedbackgabe über den Testleiter relevant.

Als möglicherweise interessanter Ansatzpunkt für den Dynamischen Test soll an dieser Stelle noch einmal kurz auf die Bedeutung der Korrekturleistung in der Feedbackbedingung Knowledge of Result eingegangen werden. Dieser Gedanke wurde schon im Rahmen der Diskussion der Ergebnisse des ersten Experiments angerissen (vgl. Abschnitt 7.1). Er bezieht sich darauf, dass die Korrekturleistung (also die Leistung in den Zweitversuchen) in der Bedingung Knowledge of Result bei etwa 45 % im ersten

Experiment und bei 40 % im zweiten Experiment lag. Die Frage ist, inwieweit diese Korrekturleistung eine über die Ratewahrscheinlichkeit beim wiederholten Beantworten einer Testfrage mit verbliebenen vier Antwortalternativen hinausgehende Leistung darstellt.

Eine Überprüfung ist zumindest mit dem Design der vorliegenden Untersuchung nicht zu erbringen und gestaltet sich auch darüber hinaus schwer. Denn eine Kontrastbedingung, die den Nutzen von Knowledge of Result auf die Korrekturleistung abschätzen lässt, ist schwer vorstellbar und nicht trivial: ein Feedback mit weniger Inhalt als Knowledge of Result gibt es nicht und Maßnahmen, die nicht genuiner Feedbacknatur sind, aber dennoch ein wiederholtes Antworten erforderlich machen ließen (z.B. *Judgement-of-Knowledge* Aufgaben), beeinflussen auf ihre Weise die Wahrnehmung und den Aufmerksamkeitsfokus des Lerners beim Bearbeiten der Aufgaben. Somit stellten sie ebenfalls eine Intervention dar. Dennoch erscheint die Frage, inwieweit eine Fehlerrückmeldung inklusive zweitem Antwortversuch oder auch nur das Einräumen einer zweiten Antwortmöglichkeit die Leistung verändert, interessant – auch in Hinblick auf Dynamische Tests.

Ausblick

Die wesentlichen nächsten Schritte zur Entwicklung eines Dynamischen Kurzzeitlerntests wurden bereits an anderen Stellen diskutiert. Sie beziehen sich vor allem auf die Überprüfung alternativer Testbedingungen, die den positiven Effekt der Inferenzprompts-via-Testleiter ebenfalls erbringen können. Zudem empfiehlt es sich, die Effektivität weiterer elaborierter Feedbacks, in erster Linie die Fehlererklärung und der metakognitive Prompt, experimentell zu überprüfen.

Neben dem Bezug auf den Dynamischen Kurzzeitlerntest, dessen Entwicklung der Ausgangspunkt für diese Arbeit war (vgl. Abschnitt 1), ergeben sich aus der Arbeit aber auch weitere Fragestellungen zur Wirkung von Feedbacks an sich. Ein wesentlicher Teil der Ergebnisse dieser Arbeit bezieht sich auf wirkungslos gebliebene Feedbackinterventionen. Dass deren Ursache sehr wahrscheinlich in einer unzureichenden Anstrengungsmotivation der Probanden zu suchen ist, wurde an mehreren Stellen diskutiert. Um das Verständnis von Feedback, den Bedingungen seiner Wirkung und eben seiner Wirkungslosigkeit, voranzubringen, erscheint es wichtig, diesen

Prozessen und Bedingungen nachzugehen. Welche Umgebungsbedingungen sind für Feedback lernförderlich und welche stellen Lerner vor Schwierigkeiten? Neben situativen Faktoren müssen hier auch stärker kognitive und vor allem motivational-emotionale Personmerkmale in Betracht gezogen werden. Die Einstellungen und Erwartungen der Leser an die Veränderbarkeit ihres Textverständnisses und ihrer Lesekompetenz erscheinen für die Förderung der Lesekompetenz besonders relevant.

Des Weiteren erscheint die Frage wichtig, inwiefern die Wahrscheinlichkeit der Fehlerkorrektur mittels Feedback von der Schwierigkeit des Items abhängt. Dass nicht jeder Leser unmittelbar jede beliebig schwere Aufgabe mithilfe von elaboriertem Feedback korrigieren kann, liegt auf der Hand. Sehr wahrscheinlich gibt es hier einen Spielraum, wie viel schwieriger die Aufgabe sein darf, damit der Leser sie (mithilfe des Feedbacks) bewältigen kann. Die Wirksamkeit von Feedbacks könnte sich nicht nur anhand der Menge der durch sie korrigierbaren Antworten bestimmen lassen, sondern möglicherweise auch daran, wie viel schwierigere Aufgaben sie zu bewältigen helfen. Erkenntnisse in dieser Hinsicht könnten beispielsweise für adaptive Test-/Interventionsverfahren genutzt werden.

Aus der Perspektive des Textverstehens erscheint es zudem interessant, wie die Leser die Informationen der Inferenzprompts, aber auch anderer (elaborierter) Feedbackarten zur Korrektur oder zum Aufbau ihrer Bedeutungsrepräsentation nutzen. Welche Aktivitäten führen sie in Folge der Rückmeldungen aus, welchen Einfluss hat ihr Strategiewissen nicht nur auf die unmittelbare Beantwortung von Verständnisfragen, sondern auch auf die Nutzung der Rückmeldungen.

14 Abbildungsverzeichnis

15 Tabellenverzeichnis

16 Literaturverzeichnis

Adams, R. (2002). Scaling PISA cognitive data. In R. Adams & M. Wu (Hrsg.), *PISA 2000 technical report* (S. 99-108). Paris: OECD.

Ainsworth, S. & Loizou, A. T. (2003). The effects of self-explaining when learning with text or diagrams. *Cognitive Science, 27*(4), 669-681.

Alber-Morgan, S. R., Matheson Ramp, E., Anderson, L. L. & Martin, C. M. (2007). Effects of repeated readings, error correction, and performance feedback on the fluency and comprehension of middle school students with behavior problems. *Journal of Special Education, 40*(1), 17-30.

Albrecht, J. E. & O'Brien, E. J. (1993). Updating a mental model: Maintaining both local and global coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(5), 1061-1070.

Allington, R. L. & McGill-Franzen, A. (2009). Comprehension difficulties among struggling readers. In S. E. Israel & G. G. Duffy (Hrsg.), *Handbook of research on reading comprehension* (S. 551-568). New York, NY: Routledge.

Anderson, R. C., Kulhavy, R. W. & Andre, T. (1971). Feedback procedures in programmed instruction. *Journal of Educational Psychology, 62*(2), 148-156.

Artelt, C. (2000). *Strategisches Lernen*. Münster: Waxmann.

Artelt, C., McElvany, N., Christmann, U., Richter, T., Groeben, N., Köster, J.et al. (2005). *Expertise - Förderung von Lesekompetenz* (Bildungsreform Band 17). Bonn: Bundesministerium für Bildung und Forschung.

Artelt, C., Schiefele, U. & Schneider, W. (2001). Predictors of reading literacy. *European Journal of Psychology of Education, 16*(3), 363-383.

Artelt, C., Stanat, P., Schneider, W. & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, K.-J. Tillmann & M. Weiß (Hrsg.), *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 69-137). Opladen: Leske + Budrich.

Artelt, C., Stanat, P., Schneider, W., Schiefele, U. & Lehmann, R. (2004). Die PISA-Studie zur Lesekompetenz: Überblick und weiterführende Analyse. In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz* (S. 139-167). Wiesbaden: VS Verlag für Sozialwissenschaften.

Ashford, S. J. & Cummings, L. L. (1983). Feedback as an individual resource: Personal strategies of creating information. *Organizational Behavior and Human Performance, 32*(3), 370-398.

Auer, M., Gruber, G., Mayringer, H. & Wimmer, H. (2005). *Salzburger Lese-Screening für die Klassenstufen 5-8 (SLS 5-8).* Göttingen: Hogrefe.

Azevedo, R. (2005a). Computer environments as metacognitive tools for enhancing learning. *Educational Psychologist, 40*(4), 193-197.

Azevedo, R. (2005b). Using hypermedia as a metacognitive tool for enhancing student learning?: The role of self-regulated learning. *Educational Psychologist, 40*(4), 199-209.

Azevedo, R. (2007). The effect of a human agent's external regulation upon college students' hypermedia learning. *Metacognition Learning, 2*(2-3), 67-87.

Azevedo, R., Cromley, J. G. & Seibert, D. (2004). Does adaptive scaffolding facilitate students' ability to regulate their learning with hypermedia? *Contemporary Educational Psychology, 29*(3), 344-370.

Azevedo, R. & Jacobson, M. J. (2008). Advances in scaffolding learning with hypertext and hypermedia: A summary and critical analysis. *Educational Technology Research and Development, 56*(1), 93-100.

Ball, L. J., Hoyle, A. M. & Towse, A. S. (2010). The facilitatory effect of negative feedback on the emergence of analogical reasoning abilities. *British Journal of Developmental Psychology, 28*(3), 583-602.

Balota, D. A., Yap, M. J. & Cortese, M. J. (2006). Visual word recognition: The journey from features to meaning (a travel update). In M. J. Traxler & M. A. Gernsbacher (Hrsg.), *Handbook of Psycholinguistics* (S. 285-375). Amsterdam: Elsevier.

Balzer, W. K., Doherty, M. E. & O'Connor, R. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin, 106*(3), 410-433.

Bandura, A. (1977). *Social learning theory*. Englewood, NJ: Prentice Hall.

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A. & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213-238.

Bannert, M. (2003). Effekte metakognitiver Lernhilfen auf den Wissenserwerb in vernetzten Lernumgebungen. *Zeitschrift für Pädagogische Psychologie, 17*(1), 13-25.

Bannert, M. (2009). Promoting self-regulated learning through prompts. *Zeitschrift für Pädagogische Psychologie, 23*(2), 139-145.

Bannert, M. & Mengelkamp, C. (2008). Assessment of metacognitive skills by means of instruction to think aloud and reflect when prompted: Does the verbalisation method affect learning? *Metacognitive Learning, 3*(1), 39-58.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*(4), 577-660.

Bartholomé, T., Stahl, E., Pieschl, S. & Bromme, R. (2006). What matters in help-seeking? A study of help effectiveness and learner-related factors. *Computers in Human Behavior, 22*(1), 113-129.

Baumert, J., Gruehn, S., Heyn, S., Köller, O. & Schnabel, K.-U. (1997). *Bildungsverläufe und psychosoziale Entwicklung im Jugendalter (BIJU)*. Berlin: Max-Planck-Institut für Bildungsforschung.

Beckmann, J. & Heckhausen, H. (2006). Motivation durch Erwartung und Anreiz. In J. Heckhausen & H. Heckhausen (Hrsg.), *Motivation und Handeln* (S. 105-142). Heidelberg: Springer.

Beckmann, J. F. (2001). *Zur Validierung des Konstrukts des intellektuellen Veränderungspotentials*. Berlin: Logos.

Beckmann, N., Beckmann, J. F. & Elliott, J. G. (2009). Self-confidence and performance goal orientation interactively predict performance in a reasoning test with accuracy feedback. *Learning and Individual Differences, 19*(2), 277-282.

Berthold, K., Nückles, M. & Renkl, A. (2007). Do learning protocols support learning strategies and outcomes? The role of cognitive and metacognitive prompts. *Learning and Instruction, 17*(5), 564-577.

Best, R. M., Rowe, M., Ozuru, Y. & McNamara, D. S. (2005). Deep-level comprehension of science texts. *Topics in Language Disorders, 25*(1), 65-83.

Blanc, N. & Tapiero, I. (2001). Updating spatial situation models: Effects of prior knowledge and task demands. *Discourse Processes, 31*(3), 241-262.

Bodemer, D., Ploetzner, R., Feuerlein, I. & Spada, H. (2004). The active integration of information during learning with dynamic and interactive visualisations. *Learning and Instruction, 14*(3), 325-341.

Boyer, K. E., Phillips, R., Ingram, A., Ha, E. Y., Wallis, M., Vouk, M.et al. (2011). Investigating the relationship between dialogue structure and tutoring effectiveness: a hidden Markov modeling approach. *International Journal of Artificial Intelligence in Education, 21*(1), 65-81.

Brehmer, B. (1979). Effect of practice on utilization of nonlinear rules in inference tasks. *Scandinavian Journal of Psychology, 20*(3), 141-149.

Brown, A. L. (1984). Metakognition, Handlungskontrolle, Selbststeuerung und andere, noch geheimnisvollere Mechanismen. In F. E. Weinert & R. H. Kluwe (Hrsg.), *Metakognition, Motivation und Lernen* (S. 60-109). Stuttgart: Kohlhammer.

Brünken, R., Seufert, T. & Zander, S. (2005). Förderung der Kohärenzbildung beim Lernen mit multiplen Repräsentationen. *Zeitschrift für Pädagogische Psychologie, 19*(1/2), 61-75.

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion*. München: Pearson.

Bußmann, H. (Hrsg.). (2002). *Lexikon der Sprachwissenschaft* (3. Aufl.). Stuttgart: Kröner.

Butler, D. L. & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245-281.

Buzhardt, J. & Semb, G. B. (2002). Item-by-item versus end-of-test feedback in a computer-based PSI course. *Journal of Behavioral Education, 11*(2), 89-104.

Cain, K. (1999). Ways of reading: How knowledge and use of strategies are related to reading comprehension. *British Journal of Developmental Psychology, 17*(2), 293-309.

Cain, K. & Oakhill, J. V. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and Writing, 11*(5-6), 489-503.

Cain, K. & Oakhill, J. V. (2007a). Cognitive bases of children's language comprehension difficulties: Where do we go from here? In K. Cain & J. Oakhill (Hrsg.), *Children's comprehension problems in oral and written language: A cognitive perspective* (S. 283-295). New York, NY: Guilford.

Cain, K. & Oakhill, J. V. (2007b). Reading comprehension difficulties: Correlates, causes, and consequences. In K. Cain & J. Oakhill (Hrsg.), *Children's comprehension problems in oral and written language: A cognitive perspective* (S. 41-75). New York, NY: Guilford.

Cain, K., Oakhill, J. V., Barnes, M. A. & Bryant, P. E. (2001). Comprehension skill, inference-making ability, and their relation to knowledge. *Memory & Cognition, 29*(6), 850-859.

Carpenter, P. A., Miyake, A. & Just, M. A. (1995). Language comprehension: Sentence and discourse processing. *Annual Review of Psychology, 46*(1), 91-120.

Casteel, M. A. (1993). Effects of inference necessity and reading goal on children's inferential generation. *Developmental Psychology, 29*(2), 346-357.

Cervone, D. & Wood, R. (1995). Goals, feedback, and the differential influence of self-regulatory processes on cognitively complex performance. *Cognitive Therapy and Research, 19*(5), 519-545.

Chi, M. T. H. (1996). Constructing self-explanations and scaffolded explanations in tutoring. *Applied Cognitive Psychology, 10*(7), 33-49.

Chi, M. T. H., Leeuw, N. d., Chiu, M.-H. & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*(3), 439-477.

Christmann, U. (1989). *Modelle der Textverarbeitung: Textbeschreibung als Textverstehen*. Münster: Aschendorff.

Christmann, U. (2002). Methoden der Verstehens- und Verständlichkeitserhebung. *Zeitschrift für Literaturwissenschaft und Linguistik, 128*, 76-97.

Christmann, U. & Groeben, N. (1999). Psychologie des Lesens. In B. Franzmann, K. Hasemann, D. Löffler & E. Schön (Hrsg.), *Handbuch Lesen* (S. 145-223). München: Saur.

Christmann, U. & Schreier, M. (2003). Kognitionspsychologie der Texverarbeitung und Konsequenzen für die Bedeutungskonstitution literarischer Texte. In F. Jannidis, G. Lauer, M. Martinez & S. Winko (Hrsg.), *Regeln der Bedeutung. Zur Theorie der Bedeutung literarischer Texte* (S. 246-285). Berlin: Walter de Gruyter.

Clariana, R. B. (1990). A comparison of answer until correct feedback and knowledge of correct response feedback under two conditions of contextualization. *Journal of Computer-Based Instruction, 17*(4), 125-129.

Clark, R. E., Howard, K. & Early, S. (2006). Motivational challenges experienced in highly complex learning environments. In J. Elen & R. E. Clark (Hrsg.), *Handling complexity in learning environments: Theory and research* (S. 27-41). Amsterdam: Elsevier.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Comer, C. L. (2007). *Benefits of the task for the delivery of negative feedback.* Manhattan, KS: Kansas State University.

Cook, A. E., Halleran, J. G. & O'Brien, E. J. (1998). What is readily available during reading? A memory-based view of text processing. *Discourse Processes, 26*(2-3), 109-129.

Corbett, A. T. & Anderson, J. R. (2001). Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. *CHI, 3*(1), 245-252.

Coté, N., Goldman, S. R. & Saul, E. U. (1998). Students making sense of informational text: Relations between processing and representation. *Discourse Processes, 25*(1), 1-53.

Davis, E. A. (2003). Prompting middle school science students for productive reflection: Generic and directed prompts. *Journal of the Learning Sciences, 12*(1), 91-142.

Davis, W., Carson, C., Ammeter, A. P. & Treadway, D. C. (2005). The interactive effects of goal orientation and feedback specificity on task performance. *Human Performance, 18*(4), 409-426.

Dempsey, J. V., Driscoll, M. P. & Litchfield, B. C. (1993). Feedback, retention, discrimination error, and feedback study time. *Journal of Research on Computing in Education, 25*(3), 303-326.

Dempsey, J. V., Driscoll, M. P. & Swindell, L. K. (1993). Text-based feedback. In J. V. Dempsey & G. C. Sales (Hrsg.), *Interactive instruction and feedback* (S. 21-54). Englewood, NJ: Educational Technology Publications.

Dempsey, J. V. & Wager, S. U. (1988). A taxonomy for the timing of feedback in computer-based instruction. *Educational Technology, 28*(10), 20-25.

Dihoff, R. E., Brosvic, G. M. & Epstein, M. L. (2003). The role of feedback during academic testing: The delay retention effect revisited. *Psychological Record, 53*(4), 533-548.

Dihoff, R. E., Brosvic, G. M., Epstein, M. L. & Cook, M. J. (2004). Provision of feedback during preparation for academic testing: Learning is enhanced by immediate but not delayed feedback. *Psychological Record, 54*(2), 207-231.

Dillon, R. F. (1997). Dynamic Testing. In R. F. Dillon (Hrsg.), *Handbook on testing* (S. 164-186). Westport: Greenwood Press.

Dörfler, T., Golke, S. & Artelt, C. (2009). Dynamic assessment and its potential for the assessment of reading. *Studies in Educational Evaluation, 35*(2-3), 77-82.

Dörfler, T., Golke, S. & Artelt, C. (2010). Dynamisches Testen der Lesekompetenz: Theoretische Grundlagen, Konzeption und Testentwicklung. In E. Klieme, D. Leutner & M. Kenk (Hrsg.), *Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes. 56. Beiheft Zeitschrift für Pädagogik* (S. 154-164).

Duffy, T. M. & Jonassen, D. H. (1992). Constructivism: New implications for instructional technology. In T. M. Duffy (Hrsg.), *Constructivism and the Technology of Instruction: A conversation* (S. 1-16). Hillsdale, NJ: Erlbaum.

Ebel, R. L. (1979). *Essentials of Educational Measurement.* Englewood, NJ: Prentice-Hall.

Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist, 46*(1), 6-25.

Farmer, M. E., Klein, R. & Bryson, S. E. (1992). Computer-assisted reading: Effects of whole-word feedback on fluency and comprehension in readers with severe disabilities. *Remedial and Special Education, 13*(2), 50-60.

Foertsch, J. & Gernsbacher, M. A. (1994). In search of complete comprehension: Getting "minimalists to work". *Discourse Processes, 18*(3), 271-296.

Frey, D., Stahlberg, D. & Fries, A. (1986). Information seeking of high- and low-anxiety subjects after receiving positive and negative self-relevant feedback. *Journal of Personality, 54*(4), 694-703.

Garrod, S., O'Brien, E. J., Morris, R. K. & Rayner, K. (1990). Elaborative inferencing as an acitve or passive process. *Journal of Experimental Psychology: Learning, Memory and Cognition, 16*(2), 250-257.

Gaynor, P. (1981). The effect of feedback delay of retention of computer-based mathematical material. *Journal of Computer-Based Instruction, 8*(2), 28-34.

Gernsbacher, M. A. & Faust, M. E. (1991). The mechanism of suppression: A component of general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(2), 245-262.

Gernsbacher, M. A., Varner, K. R. & Faust, M. E. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(3), 430-445.

Glogger, I., Holzäpfel, L., Schwonke, R., Nückles, M. & Renkl, A. (2009). Activation of learning strategies in writing learning journals: The specifity of prompts matters. *Zeitschrift für Pädagogische Psychologie, 23*(2), 95-104.

Gold, A. (2007). *Lesen kann man lernen: Lesestrategien für das 5. und 6. Schuljahr*. Göttingen: Vandenhoeck & Ruprecht.

Goodman, J., Hendrickx, M. & Wood, R. E. (2004). Feedback specificy, exploration, and learning. *Journal of Applied Psychology, 89*(2), 248-262.

Graesser, A. C., Jeon, M. & Dufty, D. (2008). Agent technologies designed to facilitate interactive knowledge construction. *Discourse Processes, 45*(4), 298-322.

Graesser, A. C., Louwerse, M. M., McNamara, D. S., Olney, A., Cai, Z. & Mitchell, H. H. (2007). Inference generation and cohesion in the construction of situation models: Some connections with computational linguistics. In F. Schmalhofer & C. A. Perfetti (Hrsg.), *Higher level language processes in the brain: Inferences and comprehension processes* (S. 289-310). Mahwah, NJ: Erlbaum.

Graesser, A. C. & McNamara, D. (2010). Self-regulated learning in learning environments with pedagogical agents that interact in natural language. *Educational Psychologist, 45*(4), 234-244.

Graesser, A. C., McNamara, D. S. & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text? In A. P. Sweet & C. E. Snow (Hrsg.), *Rethinking reading comprehension* (S. 82-98). New York, NY: Guilford Publications.

Graesser, A. C., McNamara, D. S. & VanLehn, K. (2005). Scaffolding deep comprehension strategies through point&query, AutoTutor and iSTART. *Educational Psychologist, 40*(4), 225-234.

Graesser, A. C., Millis, K. K. & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology, 48*, 163-189.

Graesser, A. C., Singer, M. & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*(3), 371-395.

Graesser, A. C. & Zwaan, R. A. (1995). Inference generation and the construction of situation models. In C. A. Weaver, S. Mannes & C. R. Fletcher (Hrsg.), *Discourse Comprehension Essays in Honor of Walter Kintsch* (S. 117-139). Hillsdale, NJ: Erlbaum.

Guerin, B. (1986). Mere presence effects in humans: A review. *Journal of Experimental Social Psychology, 22*(1), 38-77.

Guthke, J. & Wiedl, K. H. (1996). *Dynamisches Testen: Zur Psychodiagnostik der intraindividuellen Variabilität.* Göttingen: Hogrefe.

Guthrie, J. T. & Wigfield, A. (2000). Engagement and motivation in reading. In M. L. Kamil, P. Mosenthal, P. D. Pearson & R. Barr (Hrsg.), *Handbook of Reading Research* (Bd. 3, S. 403-422). Mahwah, NJ: Erlbaum.

Halldorson, M. & Singer, M. (2002). Inference processes: Integrating relevant knowledge and text information. *Discourse Processes, 34*(2), 145-161.

Hancock, T. E., Thurman, R. A. & Hubbard, D. C. (1995). An expanded control model for the use of instructional feedback. *Contemporary Educational Psychology, 20*(4), 410-425.

Hanna, G. S. (1976). Effects of total and partial feedback in multiple-choice testing upon learning. *Journal of Educational Research, 69*(5), 202-205.

Hansen, J. B. (1974). Effects of feedback, learner control, and cognitive abilities on state anxiety and performance in a computer-assisted instruction task. *Journal of Educational Psychology, 66*(2), 247-254.

Hattie, J. (1999). *Influences on student learning.* Auckland: University of Auckland.

Hattie, J. (2009). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement.* New York, NY: Routledge.

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112.

Heller, K. A. & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision (KFT 4-12+R)*. Göttingen: Beltz.

Hodapp, V. (1991). Das Prüfungsängstlichkeitsinventar TAI-G: Eine erweiterte und modifizierte Version mit vier Komponenten. *Zeitschrift für Pädagogische Psychologie, 5*(2), 121-130.

Ilgen, D. R., Fisher, C. D. & Susan, T. M. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology, 64*(4), 349-371.

Irmen, L. (2006). Satzverstehen. In J. Funke & P. A. Frensch (Hrsg.), *Handbuch der Allgemeinen Psychologie - Kognition* (S. 601-611). Göttingen: Hofgrefe.

Jacoby, J., Troutman, T., Mazursky, D. & Kuss, A. (1984). When feedback is ignored: Disutility of outcome feedback. *Journal of Applied Psychology, 69*(3), 531-545.

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

Johnson-Laird, P. N. (1989). Mental models. In M. I. Posner (Hrsg.), *Foundations of cognitive science* (S. 469-499). Cambridge, MA: MIT Press.

Johnson-Laird, P. N., Herrmann, D. & Chaffin, R. (1984). Only connections: A critique of semantic networks. *Psychological Bulletin, 96*(2), 292-315.

Jonassen, D. H. (1991). Objectivism versus constructivism: Do we need a new philosophical paradigm? *Educational Technology Research and Development, 39*(3), 5-14.

Just, M. A. & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*(1), 122-149.

Karabenick, S. A. & Knapp, J. R. (1988). Effects of Computer Privacy on Help-Seeking. *Journal of Applied Social Psychology, 18*(6), 461-472.

King, A. (2007). Beyond literal comprehension: A strategy to promote deep understanding of text. In D. S. McNamara (Hrsg.), *Reading comprehension strategies* (S. 267-290). Mahwah, NJ: Erlbaum.

Kintsch, W. (1974). *The representation of meaning in memory.* Hillsdale, NJ: Erlbaum.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*(2), 163-182.

Kintsch, W. (1994). Kognitionspsychologische Modelle des Textverstehens: Literarische Texte. In K. Reusser & M. Reusser-Weyeneth (Hrsg.), *Verstehen - Psychologischer Prozeß und didaktische Aufgabe* (S. 39-53). Bern: Hans Huber Verlag.

Kintsch, W. (1998). *Comprehension. A paradigm for cognition.* Cambridge, UK: University Press.

Kintsch, W. (2009). Learning and constructivism. In S. Tobias & T. M. Duffy (Hrsg.), *Constructivist instruction: Success or failure?* (S. 223-241). New York, NY: Routledge.

Kintsch, W. & Bates, E. (1977). Recognition memory for statements from a classroom lecture. *Journal of Experimental Psychology: Human Learning and Memory, 3*(2), 150-159.

Kintsch, W. & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363-394.

Kintsch, W., Welsch, D., Schmalhofer, F. & Zimny, S. (1990). Sentence Memory: A theoretical analysis. *Journal of Memory and Language, 29*(2), 133-159.

Klauer, K. J. (1988). Teaching for learning-to-learn: a critical appraisal with some proposals. *Instructional science, 17*(4), 351-367.

Kluger, A. N. & Adler, S. (1993). Person-versus computer-mediated feedback. *Computers in Human Behavior, 9*(1), 1-16.

Kluger, A. N. & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254-284.

Kluger, A. N. & DeNisi, A. (1998). Feedback interventions: Toward the understanding of a double-edged sword. *Current Directions in Psychological Science, 7*(3), 67-72.

Köller, O. & Baumert, J. (1998). Ein deutsches Instrument zur Erfassung von Zielorientierungen bei Schülerinnen und Schülern. *Diagnostica, 44*(4), 173-181.

Köller, O. & Möller, J. (2006). Selbstwirksamkeit. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 693-699). Weinheim: Beltz PVU.

Korsgaard, M. A. & Diddams, M. (1996). The effect of process feedback and task complexity on personal goals, information searching, and performance improvement. *Journal of Applied Social Psychology, 26*(21), 1889-1911.

Krause, U.-M. (2007). *Feedback und kooperatives Lernen.* Münster: Waxmann.

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research, 47*(1), 211-222.

Kulhavy, R. W. & Anderson, R. C. (1972). Delay-retention effect with multiple-choice tests. *Journal of Educational Psychology, 63*(5), 505-512.

Kulhavy, R. W. & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1*(4), 279-308.

Kulhavy, R. W. & Wager, W. (1993). Feedback in programmed instruction: Historical context and implications for practice. In J. Dempsey & G. Sales (Hrsg.), *Interactive instruction and feedback* (S. 3-20). Englewood, NJ: Educational Technology.

Kulhavy, R. W., White, M. T., Topp, B. W., Chan, A. L. & Adams, J. (1985). Feedback complexity and corrective efficiency. *Contemporary Educational Psychology, 10*(3), 285-291.

Kulik, J. A. & Kulik, C.-L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58*(1), 79-97.

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M.et al. (2002). *PISA 2000: Dokumentation der Erhebungsinstrumente.* Berlin: Max-Planck-Institut für Bildungsforschung.

Langer, P., Keenan, V. & Schreiner, M. E. (1995). Assisting text processing: What do we need to consider. *Psychological Reports, 76*(3), 835-845.

Latham, G. P. & Locke, E. A. (1991). Self-Regulation through goal setting. *Organizational Behavior and Human Decision Processes, 50*(2), 212-247.

Lee, H. W., Lim, K. Y. & Grabowski, B. L. (2009). Generative learning strategies and metacognitive feedback to facilitate comprehension of complex science topics and self-regulation. *Journal of Educational Multimedia and Hypermedia, 18*(1), 5-25.

Lenhard, W., Baier, H., Hoffmann, J. & Schneider, W. (2007). Automatische Bewertung offener Antworten mittels Latenter Semantischer Analyse. *Diagnostica, 53*(3), 155-165.

Lin, X. & Lehman, J. D. (1999). Supporting learning of variable control in a computer-based biology environment: Effects of prompting college students to reflect on their own thinking. *Journal of Research in Science Teaching, 36*(7), 837-858.

Long, D. L. (1994). The effects of pragmatics and discourse style on recognition memory for sentences. *Discourse Processes, 17*(2), 213-234.

Long, D. L., Graesser, A. C. & Golding, J. M. (1992). A test of the on-line status of goal-related inferences. *Journal of Memory and Language, 31*(5), 634-647.

Long, D. L., Oppy, B. J. & Seely, M. R. (1997). Individual differences in readers' sentence- and text-level representations. *Journal of Memory and Language, 36*(1), 129-145.

Long, D. L., Seely, M. R., Oppy, B. J. & Golding, J. M. (1996). The role of inferential processing in reading ability. In B. K. Britton & A. C. Graesser (Hrsg.), *Models of understanding text* (S. 189-214). Mahwah, NJ: Erlbaum.

Lussier, C. M. & Swanson, H. L. (2005). Dynamic assessment: a selective synthesis of the experimental literature. In G. M. van der Aalsvoort, W. C. M. Resing & A. J. J. M. Ruijssenaars (Hrsg.), *Learning potential assessment and cognitive training: actual research and perspectives in theory building and methodology* (S. 65-87). New York, NY: Elsevier.

Lysakowski, R. S. & Walberg, H. J. (1982). Instructional effects of cues, participation, and corrective feedback: A quantitive synthesis. *American Educational Research Journal, 19*(4), 559-578.

Mason, B. J. & Bruning, R. (2001). Providing feedback in computer-based instruction: What the research tells us. (http://dwb.unl.edu/Edit/MB/MasonBruning.html; abgerufen 02/25/2009)

Mateos, M. & Alonso, J. (1991). Metacognition and reading comprehension: Strategies for comprehension monitoring training. In M. Carretero, M. Pope, R. J. Simons & J. I. Pozo (Hrsg.), *Learning and instruction: European research in an international context* (Bd. 3, S. 273-292). Oxford, UK: Pergamon.

Mathan, S. A. & Koedinger, K. R. (2002). *An empirical assessment of comprehension fostering features in an intelligent tutoring system.* Paper presented at the 6th International Conference Intelligent Tutoring Systems, Biarritz, France and San Sebastian, Spain.

McClelland, J. L. & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1: An account of basic findings. *Psychological Review, 88*(5), 375-407.

McElvany, N., Kortenbruck, M. & Becker, M. (2008). Lesekompetenz und Lesemotivation. Entwicklung und Mediation des Zusammenhangs durch Leseverhalten. *Zeitschrift für Pädagogische Psychologie, 22*(3-4), 207-219.

McKendree, J. (1990). Effective feedback content for tutoring complex skills. *Human Computer Interaction, 5*(4), 381-413.

McKoon, G. & Ratcliff, R. (1992). Inference during reading. *Psychological Review, 99*(3), 440-466.

McNamara, D. S., Ozuru, Y., Best, R. M. & O'Reilly, T. (1995). The 4-pronged comprehension strategy framework. In D. S. McNamara (Hrsg.), *Reading Comprehension strategies* (S. 465-496). New York, NY: Erlbaum.

Meyer, B. J. F., Wijekumar, K., Middlemiss, W., Higley, K., Lei, P.-W., Meier, C.et al. (2010). Web-based tutoring of the structure strategy with or without elaborated feedback or choice for fifth- and seventh-grade readers. *Reading Research Quarterly, 45*(1), 62-92.

Möller, J. & Schiefele, U. (2004). Motivationale Grundlagen der Lesekompetenz. In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz* (S. 101-124). Wiesbaden: VS Verlag für Sozialwissenschaften.

Moos, D. C. & Azevedo, R. (2008). Monitoring, planning, and self-efficacy during learning with hypermedia: The impact of conceptual scaffolds. *Computers in Human Behavior, 24*(4), 1686-1706.

Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional Science, 32*(1-2), 99-113.

Morris, C. D. & Bransford, J. D. (1982). Effective elaboration and inferential reasoning. *Memory & Cognition, 10*(2), 188-193.

Morrison, G. R., Ross, S. M., Gopalakrishnan, M. & Casey, J. (1995). The effects of feedback and incentives on achievement in computer-based instruction. *Contemporary Educational Psychology, 20*(1), 32-50.

Morrow, D. G. & Bower, G. H. (1989). Updating situation models during narrative comprehension. *Journal of Memory and Language, 28*(3), 292-312.

Mory, E. H. (2004). Feedback research revisited. In D. H. Jonassen (Hrsg.), *Handbook of research on educational communications and technology* (2. Aufl., S. 745-783). Mahwah, NJ: Erlbaum.

Murphy, G. L. & Shapiro, A. M. (1994). Forgetting of verbatim information in discourse. *Memory & Cognition, 22*(1), 85-94.

Murphy, P. (2010). Web-based collaborative reading exercises for learners in remote locations: The effects of computer-mediated feedback and interaction via computer-mediated communication. *ReCALL, 22*(2), 112-134.

Musch, J. (1999). Die Gestaltung von Feedback in computergestützten Lernumgebungen: Modelle und Befunde. *Zeitschrift für Pädagogische Psychologie, 13*(3), 148-160.

Myers, M. & Paris, S. G. (1978). Children's metacognitive knowledge about reading. *Journal of Educational Psychology, 70*(5), 680-690.

Narciss, S. (2006). *Informatives tutorielles Feedback*. Münster: Waxmann.

Narciss, S. & Huth, K. (2004). How to design informative tutoring feedback for multimedia learning. In H. Niegemann, R. Brünken & D. Leutner (Hrsg.), *Instructional design for multimedia learning* (S. 181-195). Münster: Waxmann.

Narciss, S., Koerndle, H. & Proske, A. (2006). Promoting self-regulated learning in web-based learning environments. In G. Clarebout & J. Elen (Hrsg.), *Avoiding simplicity, confronting complexity: Advances in studying and designing powerful (computer-based) learning environments* (S. 219-231). Rotterdam: Sense Publishers.

Nation, K. (2005). Children's reading comprehension difficulties. In M. J. Snowling & C. Hulme (Hrsg.), *The science of reading: A handbook* (S. 248-265). Malden, MA: Blackwell.

National Institute of Child Health and Human Development, NIH, DHHS (2000). *Report of the National Reading Panel: Teaching children to read.* Washington, DC: US Government Printing Office.

Nicholls, J. G. (1984). Achievement motivation: conceptions of ability, subjective experience, task choice, and performance. *Psychological Review, 91*(3), 328-346.

Noordman, L. G. M. & Vonk, W. (1992). Readers' knowledge and the control of inferences in reading. *Language and Cognitive Processes, 7*(3/4), 373-391.

O'Brien, E. J. & Albrecht, J. E. (1992). Comprehension strategies in the development of a mental model. *Journal of Experimental Psychology: Learning, Memory and Cognition, 18*(4), 777-784.

O'Brien, E. J., Shank, D. M., Myers, J. L. & Rayner, K. (1988). Elaborative inferences during reading: do they occur on-line? *Journal of Experimental Psychology: Learning, Memory and Cognition, 14*(3), 410-420.

Oakhill, J. V. (1982). Constructive processes in skilled and less skilled comprehenders' memory for sentences. *British Journal of Psychology, 73*(1), 13-20.

Oakhill, J. V. (1984). Inferential and memory skills in children's comprehension of stories. *British Journal of Educational Psychology, 54*(1), 31-39.

Oakhill, J. V. (1993). Children's difficulties in reading comprehension. *Educational PSychology Review, 5*(3), 223-237.

Oakhill, J. V. & Cain, K. (2007). Issues of causality in children's reading comprehension. In D. McNamara (Hrsg.), *Reading comprehension strategies: Theories, interventions, and technologies* (S. 47-71). New York, NY: Erlbaum.

Oakhill, J. V. & Garnham, A. (1988). *Becoming a skilled reader*. Oxford, UK: Blackwell.

Oakhill, J. V. & Yuill, N. (1996). Higher order factors in comprehension disability: Processes and remediation. In C. Cornoldi & J. Oakhill (Hrsg.), *Reading comprehension difficulties* (S. 69-92). Mahwah, NJ: Erlbaum.

OECD. (2005). *PISA 2003 technical report*. Paris: OECD.

Ozuru, Y., Best, R. M., Bell, C., Witherspoon, A. & McNamara, D. S. (2007). Influence of question format and text availability on the assessment of expository text comprehension. *Cognition and Instruction, 25*(4), 399-438.

Paas, F., Renkl, A. & Sweller, J. (2003). Cognitive load theory and instructional design: recent developments. *Educational Psychologist, 38*(1), 1-4.

Paas, F., Renkl, A. & Sweller, J. (2004). Cognitive load theory: instructional implications of the interaction between information structures and cognitive architecture. *Instructional Science 32*(1-2), 1-8.

Pany, D., McCoy, K. M. & Peters, E. E. (1981). Effects of corrective feedback on comprehension skills of remedial students. *Journal of Reading Behavior, 13*(2), 131-143.

Paris, S. G., Wasik, B. A. & Turner, J. C. (1991). The development of strategic readers. In R. Barr, M. L. Kamil, P. Mosenthal & P. D. Pearson (Hrsg.), *Handbook of reading research* (Bd. 2, S. 609-640). New York, NY: Longman.

Pearson, P. D. & Fielding, L. (1991). Comprehension instruction. In R. Barr, M. L. Kamil, P. Mosenthal & P. D. Pearson (Hrsg.), *Handbook of reading research* (S. 815-860). Mahwah, NJ: Erlbaum.

Peeck, J. (1979). Effects of differential feedback on the answering of two types of questions by fifth- and sixth-graders. *British Journal of Educational Psychology, 49*(1), 87-92.

Perfetti, C. A. (1985). *Reading ability*. New York, NY: Oxford University Press.

Perfetti, C. A. (1994). Psycholinguistics and reading ability. In M. A. Gernsbacher (Hrsg.), *Handbook of psycholinguistics* (S. 849-894). San Diego, CA: Academic Press.

Perfetti, C. A. & Britt, M. A. (1995). Where do propositions come from? In C. A. Weaver, S. Mannes & C. R. Fletcher (Hrsg.), *Discourse comprehension: Essays in honor of Walter Kintsch* (S. 11-35). Hillsdale, NJ: Erlbaum.

Perfetti, C. A., Marron, M. A. & Foltz, P. W. (1996). Sources of comprehension failure: Theoretical perspectives and case studies. In C. Cornoldi & J. Oakhill (Hrsg.), *Reading comprehension difficulties* (S. 137-165). Mahwah, NJ: Erlbaum.

Peverly, S. T. & Wood, R. (2001). The effects of adjunct questions and feedback on improving the reading comprehension skills of learning-disabled adolescents. *Contemporary Educational Psychology, 26*(1), 25-43.

Pressley, M. (2000). What should comprehension instruction be the instruction of? In M. L. Kamil, P. Mosenthal, P. D. Pearson & R. Barr (Hrsg.), *Handbook of Reading Research* (Bd. 3, S. 545-562). Mahwah, NJ: Erlbaum.

Pressley, M., Borkowski, J. G. & Schneider, W. (1989). Good information processing: What it is and how education can promote it. *International Journal of Educational Research, 13*(8), 857-867.

Prüfer, P. & Rexroth, M. (2005). *Kognitive Interviews*. ZUMA How-to-Reihe, Nr. 15.

Puntambekar, S. & Stylianou, A. (2005). Designing navigation support in hypertext systems based on navigation patterns. *Instructional science, 33*(5-6), 451-481.

Radvansky, G. A. (2005). Situation models, propositions, and the fan effect. *Psychonomic Bulletin & Review, 12*(3), 478-483.

Rakoczy, K., Klieme, E., Bürgermeister, A. & Harks, B. (2008). The interplay between student evaluation and instruction. *Journal of Psychology, 216*(2), 111-124.

Raphael, T. E., George, M., Weber, C. M. & Nies, A. (2009). Approaches to teaching reading comprehension. In S. E. Israel & G. G. Duffy (Hrsg.), *Handbook of research on reading comprehension* (S. 449-467). New York, NY: Routledge.

Raphael, T. E. & Wonnascott, C. A. (1985). Heightening fourth-grade students' sensitivity to sources of information for answering comprehension questions. *Reading Research Quarterly, 20*(3), 282-296.

Richter, T. & Christmann, U. (2002). Lesekompetenz: Prozessebenen und interindividuelle Unterschiede. In N. Groeben & B. Hurrelmann (Hrsg.), *Lesekompetenz. Bedingungen, Dimensionen, Funktionen* (S. 25-58). Weinheim: Juventa Verlag.

Rinck, M. (2000). Situationsmodelle und das Verstehen von Erzähltexten: Befunde und Probleme. *Psychologische Rundschau, 51*(3), 115-122.

Rinck, M. (2008). Spatial situation models and narrative comprehension. In M. A. Gluck, J. R. Anderson & S. M. Kosslyn (Hrsg.), *Memory and mind: A Festschrift for Gordon H. Bower* (S. 359-370). New York, NY: Erlbaum.

Roberts, F. C. & Park, O.-C. (1984). Feedback strategies and cognitive style in computer-based instruction. *Journal of Instructional Psychology, 11*(1), 63-74.

Robin, A. L. (1978). The timing of feedback in personalized instruction. *Journal of Personalized Instruction, 3*(2), 81-88.

Rosebrock, C. & Nix, D. (2006). Forschungsüberblick: Leseflüssigkeit (Fluency) in der amerikanischen Leseforschung und -didaktik. *Didaktik Deutsch, 20*, 90-112.

Rost, D. H. & Schilling, S. (2006). Leseverständnis. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 450-460). Weinheim: Beltz PVU.

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion*. Bern: Huber.

Salomon, G. & Globerson, T. (1987). Skill may not be enough: The role of mindfulness in learning and transfer. *International Journal of Educational Research, 11*(6), 623-637.

Sanders, T. J. M., Spooren, W. P. M. & Noordman, L. G. M. (1992). Toward a taxonomy of coherence relations. *Discourse Processes, 15*(1), 1-35.

Sandoval, W. A., Trafton, J. G. & Reiser, B. J. (1995). The effects of self-explanation on studying examples and solving problems. In J. D. Moore & J. F. Lehman (Hrsg.), *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (S. 253-258). Mahwah, NJ: Erlbaum.

Sanford, A. J. (2002). Context, attention and depth of processing during interpretation. *Mind & Language, 17*(1-2), 188-206.

Schimmel, B. J. (1988). Providing meaningful feedback in courseware. In D. H. Jonassen (Hrsg.), *Instructional designs for microcomputer courseware* (S. 183-195). London: Erlbaum.

Schmalhofer, F. & Glavanov, D. (1986). Three components of understanding a Programmer's Manual: Verbatim, propositional, and situational representations. *Journal of Memory and Language, 25*(3), 279-294.

Schnotz, W. (1991). Metacognition and self regulation in text processing: Some comments. In M. Carretero, M. Pope, R. Simons & J. I. Pozo (Hrsg.), *Learning and instruction: European research in an international context*. Vol. III (S. 365-375). Oxford: Pergamon.

Schnotz, W. (2006). Was geschieht im Kopf des Lesers? Mentale Konstruktionsprozesse beim Textverstehen aus der Sicht der Psychologie und der kognitiven Linguistik. In H. Blühdorn, E. Breindl & U. H. Waßner (Hrsg.), *Text-Verstehen. Grammatik und darüber hinaus* (S. 222-238). Berlin: Walter de Gruyter.

Schunk, D. H. & Rice, J. M. (1986). Extended attributional feedback: Sequence effects during remedial reading instruction. *Journal of Early Adolescence, 6*(1), 55-66.

Schunk, D. H. & Rice, J. M. (1991). Learning goals and progress feedback during reading comprehension instruction. *Journal of Reading Behavior, 23*(3), 351-364.

Schunk, D. H. & Rice, J. M. (1993). Strategy fading and progress feedback: Effects on self-efficacy and comprehension among students receiving remedial reading services. *Journal of Special Education, 27*(3), 257-276.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189.

Singer, M. (1994). Discourse inference processes. In M. A. Gernsbacher (Hrsg.), *Handbook of Psycholinguistics* (S. 479-515). San Diego, CA: Academic Press.

Singer, M., Graesser, A. C. & Trabasso, T. (1994). Minimal or global inference during reading. *Journal of Memory and Language, 33*(4), 421-441.

Singer, M., Harkness, D. & Stewart, S. T. (1997). Constructing inferences in expository text comprehension. *Discourse Processes, 24*(2-3), 199-228.

Singer, M. & Leon, J. (2007). Psychological studies of higher language processes: Behavioral and empirical approaches. In F. Schmalhofer & C. A. Perfetti (Hrsg.), *Higher level language processes in the brain: Inferences and comprehension processes* (S. 9-25). Mahwah, NJ: Erlbaum.

Smith, T. A. & Kimball, D. R. (2010). Learning from feedback: Spacing and the delay–retention effect. *Journal of Experimental Psychology: Learning, Memory and Cognition, 36*(1), 80-95.

Snow, R. E. (1997). Individual differences. In R. D. Tennyson, F. Schott, N. M. Seel & S. Dijkstra (Hrsg.), *Instructional design: Volume I: Theory, research, and models* (S. 215-241). Mahwah, NJ: Erlbaum.

Solman, R. T. & Wu, H.-M. (1995). Pictures as feedback in single word learning. *Educational Psychology, 15*(3), 227-244.

Song, S. H. & Keller, J. M. (2001). Effectiveness of motivationally adaptive computer-assisted instruction on the dynamic aspects of motivation. *Educational Technology Research and Development, 49*(2), 5-22.

Souvignier, E. & Antoniou, F. (2007). Förderung des Leseverständnisses bei Schülerinnen und Schülern mit Lernschwierigkeiten - eine Metaanalyse. *Vierteljahresschrift für Heilpädagogik und ihre Nachbargebiete, 76*(1), 46-62.

Spielberger, C. D. & Vagg, P. R. (1995). Test anxiety: A transactional process model. In C. D. Spielberger & P. R. Vagg (Hrsg.), *Test anxiety: Theory, assessment, and treatment* (S. 3-14). Philadelphia, PA: Taylor & Francis.

Spooner, A. L. R., Gathercole, S. E. & Baddeley, A. D. (2006). Does weak reading comprehension reflect an integration deficit? *Journal of Research in Reading, 29*(2), 173-193.

Stake, J. E. (1982). Reactions to positive and negative feedback: Enhancement and consistency effects. *Social Behavior and Personality, 10*(2), 151-156.

Stark, R., Tyroller, M., Krause, U.-M. & Mandl, H. (2008). Effekte einer metakognitiven Promptingmaßnahme beim situierten, beispielbasierten Lernen im Bereich Korrelationsrechnung. *Zeitschrift für Pädagogische Psychologie, 22*(1), 59-71.

Steele Johnson, D., Perlow, R. & Pieper, K. F. (1993). Differences in task performance as a function of type of feedback: learning-oriented versus performance-oriented feedback. *Journal of Applied Social Psychology, 23*(4), 303-320.

Sternberg, R. J. & Grigorenko, E. L. (2002). *Dynamic testing.* New York, NY: Cambridge University Press.

Sweller, J., van Merriënboer, J. J. G. & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251-296.

Tapiero, I. & Otero, J. (2002). Situation models as retrieval structures: Effects on the global coherence of science texts. In J. Otero, J. A. Leon & A. C. Graesser (Hrsg.), *The psychology of science text comprehension* (S. 179-198). Mahwah, NJ: Erlbaum.

Therriault, D. J. & Rinck, M. (2007). Multidimensional situation models. In F. Schmalhofer & C. A. Perfetti (Hrsg.), *Higher level language processes in the brain: Inference and comprehension processes* (S. 311-327). Mahwah, NJ: Erlbaum.

Therriault, D. J., Rinck, M. & Zwaan, R. A. (2006). Assessing the influence of dimensional focus during situation model construction. *Memory & Cognition, 34*(1), 78-89.

Thillmann, H., Künsting, J., Wirth, J. & Leutner, D. (2009). Is it merely a question of "What" to prompt or also "When" to prompt?: The role of point of presentation time of prompts in self-regulated learning. *Zeitschrift für Pädagogische Psychologie, 23*(2), 105-115.

Thompson, W. B. (1998). Metamemory accuracy: Effects of feedback and the stability of individual differences. *American Journal of Psychology, 111*(1), 33-42.

Thorndike, E. L. (1932). *The fundamentals of learning*. New York, NY: Teachers College Press.

Trabasso, T. & Magliano, J. P. (1996). Conscious understanding during comprehension. *Discourse Processes, 21*(3), 255-287.

Trabasso, T., Suh, S., Payton, P. & Jain, R. (1995). Explanatory inferences and other strategies during comprehension and their effect on recall. In R. F. J. Lorch & E. J. O'Brien (Hrsg.), *Sources of Coherence in Reading* (S. 219-239). Hillsdale, NJ: Erlbaum.

Trabasso, T. & Wiley, J. (2005). Goal plans of action and inferences during comprehension of narratives. *Discourse Processes, 39*(2-3), 129-164.

Tyroller, M. (2005). *Effekte metakognitiver Prompts beim computerbasierten Statistiklernen.* München: Ludwig-Maximilians-Universität München.

van den Boom, G., Paas, F. & van Merrienboer, J. J. G. (2007). Effects of elicited reflections combined with tutor or peer feedback on self-regulated learning and learning outcomes. *Learning and Instruction, 17*(5), 532-548.

van den Broek, P. (1994). Comprehension and memory of narrative texts: Inferences and coherence. In M. A. Gernsbacher (Hrsg.), *Handbook of Psycholinguistics* (S. 539-588). San Diego, CA: Academic Press.

van den Broek, P. & Kremer, K. E. (1999). The mind in action: What it means to comprehend during reading. In B. Taylor, M. Graves & P. van den Broek (Hrsg.), *Reading for Meaning* (S. 1-31). New York, NY: Teachers College Press.

van Dijk, T. A. (1980). *Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition*. Hillsdale, NJ: Erlbaum.

van Dijk, T. A. & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York, NY: Academic Press.

van Someren, M. W., Barnard, Y. F. & Sandberg, J. A. C. (1994). *The think aloud method: A practical guide to modelling cognitive processes.* London: Academic Press.

Vygotsky, L. S. (1964). *Denken und Sprechen.* Berlin: Akademie-Verlag.

Webb, J. M., Stock, W. A. & McCarthy, M. T. (1994). The effects of feedback timing on learning facts: The role of response confidence. *Contemporary Educational Psychology, 19*(3), 251-265.

Whyte, M. M., Karolick, D. M., Nielsen, M. C., Elder, G. D. & Hawley, W. T. (1995). Cognitive styles and feedback in computer-assisted instruction. *Journal of Educational Computing Research, 12*(2), 195-203.

Wichmann, A. & Leutner, D. (2009). Inquiry learning: Multilevel support with respect to inquiry, explanations and regulation during an inquiry cycle. *Zeitschrift für Pädagogische Psychologie, 23*(2), 117-127.

Winne, P. H., Graham, L. & Prock, L. (1993). A model of poor readers' text based inferencing: Effects of explanatory feedback. *Reading research quarterly, 28*(1), 53-66.

Wirth, J. (2009). Promoting self-regulated learning through prompts. *Zeitschrift für Pädagogische Psychologie, 23*(2), 91-94.

Wu, M., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). *Conquest*. Camberwell: Acer.

Yang, Y.-F., Yeh, H.-C. & Wong, W.-K. (2008). Constructing mental representation of reference by feedback in a computer system. *Computers in Human Behavior, 24*(5), 1959-1976.

Yuill, N., Oakhill, J. V. & Parkin, A. (1989). Working memory, comprehension ability and the resolution of text anomaly. *British Journal of Psychology, 80*(3), 351-361.

Zeidner, M. (2007). Test anxiety in educational contexts: concepts, findings, and future directions. In P. A. Schutz & R. Pekrun (Hrsg.), *Emotion in education* (S. 165-184). San Diego, CA: Elsevier.

Zwaan, R. A. (1994). Effect of genre expectations on text comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition, 20*(4), 920-933.

Zwaan, R. A. (1996). Toward a model of literary comprehension. In B. K. Britton & A. C. Graesser (Hrsg.), *Models of Understanding Text* (S. 241-255). Mahwah, NJ: Erlbaum.

Zwaan, R. A. (1998). Constructing multidimensional situation models during reading. *Scientific Studies of Reading, 2*(3), 199-220.

Zwaan, R. A., Magliano, J. P. & Graesser, A. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21*(2), 386-397.

Zwaan, R. A. & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*(2), 162-185.

Zwaan, R. A. & Rapp, D. N. (2006). Discourse comprehension. In M. J. Traxler & M. A. Gernsbacher (Hrsg.), *Handbook of psycholinguistics* (2. Aufl., S. 725-764). Amsterdam: Elsevier.

Zwaan, R. A. & Singer, M. (2003). Text comprehension. In A. C. Graesser, M. A. Gernsbacher & S. R. Goldman (Hrsg.), *Handbook of discourse processes* (S. 83-121). Mahwah, NJ: Erlbaum.

Anhang

Anhang A Instrumente (Fragebogen)

Anhang B Ergebnisse zentraler Analysen unter Nutzung der alternativen Cut-off-Werte für „Durchklicker“ (Experiment 1)

Anhang C Ergebnisse zentraler Analysen unter Nutzung der alternativen Cut-off-Werte für „Durchklicker“ (Experiment 2)

Anhang A:

Instrumente (Fragebogen)

LESEINTERESSE

	Items	Quelle
	„Wie ist das bei dir mit dem Lesen? Bitte kreuze an, wie sehr die folgenden Aussagen auf dich zutreffen!“	
1.	Weil mir das Lesen Spaß macht, würde ich es nicht gerne aufgeben.	PISA 2000 (Kunter et al., 2002), Skala: Interesse (Lesen)
2.	Ich lese in meiner Freizeit.	PISA 2000 (Kunter et al., 2002), Skala: Interesse (Lesen)
3.	Wenn ich lese, vergesse ich manchmal alles um mich herum.	PISA 2000 (Kunter et al., 2002), Skala: Interesse (Lesen)
4.	Lesen ist mir persönlich wichtig.	PISA 2000 (Kunter et al., 2002), Skala: Interesse (Lesen)
5.	Lesen macht mir Spaß.	McElvany et al. (2008), Skala: Lesemotivation
6.	Lesen ist für mich langweilig.	McElvany et al. (2008), Skala: Lesemotivation
7.	Lesen ist eines meiner liebsten Hobbies.	Berliner Leselängsschnitt (Lesen 3-6), Skala: Leselust (nicht veröffentlicht)
8.	Lesen macht mir großes Vergnügen.	Berliner Leselängsschnitt (Lesen 3-6), Skala: Leselust (nicht veröffentlicht)
9.	Ich lese gern.	McElvany et al. (2008), Skala: Lesemotivation

SELBSTWIRKSAMKEIT

	Items	Quelle
	„Bitte kreuze an, wie sehr die folgenden Aussagen zum Lesen auf dich zutreffen!“	
1.	Wenn ich genug übe, kann ich gut lesen.	Berliner Leselängsschnitt, Skala: Selbstwirksamkeit Lesen (nicht veröffentlicht)
2.	Wenn ich mich genug anstrenge, kann ich gut lesen.	Berliner Leselängsschnitt, Skala: Selbstwirksamkeit Lesen (nicht veröffentlicht)
3.	Wenn ich mir Mühe gebe, kann ich gut lesen.	Berliner Leselängsschnitt, Skala: Selbstwirksamkeit Lesen (nicht veröffentlicht)
4.	Ich bin sicher, dass ich auch den schwierigsten Stoff in Unterrichtstexten verstehen kann.	PISA 2000 (Kunter et al., 2002), Skala: Selfefficacy
5.	Ich bin überzeugt, dass ich auch den kompliziertesten Stoff, den der Lehrer vorstellt, verstehen kann.	PISA 2000 (Kunter et al., 2002), Skala: Selfefficacy
6.	Ich bin überzeugt, dass ich in Hausaufgaben und Klassenarbeiten gute Leistungen erzielen kann.	PISA 2000 (Kunter et al., 2002), Skala: Selfefficacy
7.	Ich bin überzeugt, dass ich die Fertigkeiten, die gelehrt	PISA 2000 (Kunter et al.,

	werden, beherrschen kann.	2002), Skala: Selfefficacy
8.	Ich bin sicher, dass ich auch die schwierigsten Texte verstehen kann.	Gold (2007; S. 43)
9.	Wenn ich mich nur lange genug mit einem Text beschäftige, dann kann ich jeden Text verstehen.	Gold (2007; S. 43), leicht adaptiert: „jeden Text verstehen“ statt „jeden Text gut verstehen“
10.	Auch lange Texte kann ich gut verstehen, wenn ich mich anstrenge.	Eigenentwicklung
11.	Wenn ich mir Mühe gebe, kann ich auch Texte über mir unbekannte Sachen gut verstehen.	Eigenentwicklung

ZIELORIENTIERUNGEN

	Items	**Quelle**
	„Bitte kreuze an, wie sehr die folgenden Aussagen auf dich zutreffen! Ich fühle mich in der Schule wirklich zufrieden, wenn…“	
1.	… ich mehr weiß als die anderen.	BIJU (Baumert et al., 1997), Skala: Soziale Vergleichs-orientierung (Ego-Orientation)
2.	… ich als einziger die richtige Antwort weiß.	BIJU (Baumert et al., 1997), Skala: Soziale Vergleichs-orientierung (Ego-Orientation)
3.	… ich bessere Noten bekomme als die anderen.	PISA 2000 (Kunter et al., 2002), Skala Ego-Orientation
4.	… ich zeigen kann, dass ich schlau bin.	PISA 2000 (Kunter et al., 2002), Skala Ego-Orientation
5.	… der Unterricht mich zum Nachdenken bringt.	BIJU (Baumert et al., 1997), Skala: Aufgabenorientierung (Task-Orientation)
6.	... das Gelernte wirklich Sinn für mich macht.	BIJU (Baumert et al., 1997), Skala: Aufgabenorientierung (Task-Orientation)
7.	… das Gelernte mich dazu bringt, mehr erfahren zu wollen.	PISA 2000 (Kunter et al., 2002), Skala Task Orientation
8.	… ich einen neuen Weg herausfinde, ein Problem zu lösen.	PISA 2000 (Kunter et al., 2002), Skala: Task Orientation
9.	… ich ein kompliziertes Problem endlich bewältigt habe.	PISA 2000 (Kunter et al., 2002), Skala: Task Orientation

TESTANGST

	Text	Quelle
	„Wenn du an die Texte und Aufgaben denkst, die du gleich bearbeiten wirst, wie sehr treffen dann die folgenden Aussagen auf dich zu?“	
1.	Ich mache mir Sorgen, ob ich auch alles schaffe.	Hodapp (1991)
2.	Ich frage mich, ob meine Leistung ausreicht.	Hodapp (1991)
3.	Ich fühle mich unwohl.	Hodapp (1991), adaptiert „unwohl“
4.	Ich denke daran, wie wichtig mir ein gutes Ergebnis ist.	Hodapp (1991)
5.	Ich mache mir Gedanken über mein Abschneiden.	Hodapp (1991)
6.	Ich fühle mich ängstlich.	Hodapp (1991)
7.	Ich denke daran, was passiert, wenn ich schlecht abschneide.	Hodapp (1991)
8.	Ich bin aufgeregt.	Hodapp (1991)

Variante, die nach dem Experiment eingesetzt wurde

	Text	Quelle
	„Wenn du an die Bearbeitung der bisherigen Texte und Aufgaben denkst, wie sehr treffen dann die folgenden Aussagen auf dich zu?“	
1.	Ich mache mir Sorgen, ob ich auch alles geschafft habe.	Hodapp (1991), Zeitform geändert
2.	Ich frage mich, ob meine Leistung ausgereicht hat.	Hodapp (1991), Zeitform geändert
3.	Ich fühle mich unwohl.	Hodapp (1991), adaptiert „unwohl“
4.	Ich denke daran, wie wichtig mir ein gutes Ergebnis ist.	Hodapp (1991)
5.	Ich mache mir Gedanken über mein Abschneiden.	Hodapp (1991)
6.	Ich fühle mich ängstlich.	Hodapp (1991)
7.	Ich denke daran, was passiert, wenn ich schlecht abgeschnitten habe.	Hodapp (1991), Zeitform geändert
8.	Ich bin aufgeregt.	Hodapp (1991)

Anhang B:

Ergebnisse zentraler Analysen unter Nutzung der alternativen Cut-off-Werte für „Durchklicker“, die für Fallausschlüsse herangezogen werden

Analysen für Experiment 1

ANOVAs für Leistung in Erstversuchen (Treatmentphase), N = 34 Items

	Cut-off-Kriterien: wann ein Item als zu schnell beantwortet gilt								
	< 7 Sekunden			< 8 Sekunden			< 9 Sekunden		
Versuchsgruppe	N	*M*	*SD*	N	*M*	*SD*	N	*M*	*SD*
Metakognitiver Prompt	108	14.47	5.50	103	14.47	5.59	101	14.60	5.55
Fehlererklärung	94	14.64	5.34	92	14.76	5.32	90	14.92	5.26
Inferenzprompt	96	14.78	5.47	95	14.82	5.48	94	14.81	5.51
Knowledge of Result	95	14.40	5.11	93	14.54	5.08	89	14.69	5.04
Kein Feedback	92	15.55	5.34	89	15.66	5.49	88	15.75	5.45
	$N_{gesamt} = 485$			$N_{gesamt} = 472$			$N_{gesamt} = 462$		
	F (4, 480) = 0.70			F (4, 467) = 0.72			F (4, 457) = 0.66		
	$p > .05$			$p > .05$			$p > .05$		
	$\eta^2 = .006$			$\eta^2 = .006$			$\eta^2 = .006$		

Anmerkungen. Analog zu den im Hauptteil der Arbeit berichteten Ergebnissen wurden für die Berechnungen zusätzlich die Fälle mit unvollständigen Datensätzen des Experiments ausgeschlossen.

ANOVAs für Leistung in Zweitantworten (Treatmentphase) als relative Lösungshäufigkeiten (relativiert an Anzahl zweiter Versuche)

	Cut-off-Kriterien: wann ein Item als zu schnell beantwortet gilt								
	< 7 Sekunden			< 8 Sekunden			< 9 Sekunden		
Versuchsgruppe	N	*M*	*SD*	N	*M*	*SD*	N	*M*	*SD*
Metakognitiver Prompt	108	0.44	0.16	103	0.44	0.16	101	0.44	0.16
Fehlererklärung	94	0.42	0.14	92	0.42	0.14	90	0.42	0.14
Inferenzprompt	96	0.42	0.1	95	0.42	0.14	94	0.42	0.14
Knowledge of Result	95	0.45	0.16	93	0.45	0.16	89	0.45	0.16
	$N_{gesamt} = 393$			$N_{gesamt} = 383$			$N_{gesamt} = 374$		
	F (3, 389) = 0.70			F (3, 379) = 0.75			F (3, 370) = 0.81		
	$p > .05$			$p > .05$			$p > .05$		
	$\eta^2 = .005$			$\eta^2 = .006$			$\eta^2 = .007$		

Anmerkungen. Analog zu den im Hauptteil der Arbeit berichteten Ergebnissen wurden für die Berechnungen zusätzlich die Fälle mit unvollständigen Datensätzen des Experiments ausgeschlossen.

ANOVAs für Leistung in Posttest, N = 13 Items

	Cut-off-Kriterien: wann ein Item als zu schnell beantwortet gilt								
	< 7 Sekunden			< 8 Sekunden			< 9 Sekunden		
Versuchsgruppe	N	*M*	*SD*	N	*M*	*SD*	N	*M*	*SD*
Metakognitiver Prompt	77	5.91	2.56	72	5.97	2.62	70	6.00	2.64
Fehlererklärung	61	5.70	2.95	58	5.81	2.99	57	5.79	3.01
Inferenzprompt	72	5.40	2.90	68	5.41	2.82	62	5.56	2.84
Knowledge of Result	69	6.13	2.65	66	6.11	2.69	59	6.19	2.68
Kein Feedback	71	6.01	2.72	65	6.28	2.61	64	6.33	2.60
	$N_{gesamt} = 350$ F (4, 345) = 0.77 $p > .05$ $\eta^2 = .009$			$N_{gesamt} = 329$ F (4, 324) = 0.97 $p > .05$ $\eta^2 = .012$			$N_{gesamt} = 312$ F (4, 307) = 0.76 $p > .05$ $\eta^2 = .010$		

Anmerkungen. Analog zu den im Hauptteil der Arbeit berichteten Analysen wurden für die hier berichteten alternativen Berechnungen zusätzlich die Fälle mit unvollständigen Datensätzen des Posttests und des Experiments sowie die Fälle, die aufgrund des geltenden Kriteriums als „Durchklicker“ im Experiment identifiziert wurden, ausgeschlossen.

ANOVAs für Leistung in Follow-up, N = 28 Items

	Cut-off-Kriterien: wann ein Item als zu schnell beantwortet gilt								
	< 7 Sekunden			< 8 Sekunden			< 9 Sekunden		
Versuchsgruppe	N	*M*	*SD*	N	*M*	*SD*	N	*M*	*SD*
Metakognitiver Prompt	93	17.34	5.40	88	17.44	5.49	86	17.66	5.35
Fehlererklärung	90	17.02	5.77	88	17.09	5.82	86	17.30	5.71
Inferenzprompt	93	17.63	5.04	92	17.72	5.00	92	17.72	5.00
Knowledge of Result	91	17.63	5.08	89	17.79	4.98	85	17.93	4.88
Kein Feedback	87	18.20	5.21	84	18.35	5.23	83	18.41	5.23
	$N_{gesamt} = 454$ F (4, 449) = 0.59 $p > .05$ $\eta^2 = .005$			$N_{gesamt} = 441$ F (4, 436) = .65 $p > .05$ $\eta^2 = .006$			$N_{gesamt} = 432$ F (4, 427) = 0.51 $p > .05$ $\eta^2 = .005$		

Anmerkungen. Für die Analysen des Follow-ups wurden die Fälle ausgeschlossen, die im Experiment (Treatmentphase) entweder aufgrund unvollständiger Datensätze und/oder anhand des geltenden Kriteriums (siehe Tabelle) als „Durchklicker“ identifiziert wurden.

ANOVAs für Testangst, Komponente Besorgtheit (5 Items), nach dem Experiment erhoben

	Cut-off-Kriterien: wann ein Item als zu schnell beantwortet gilt								
	< 7 Sekunden			< 8 Sekunden			< 9 Sekunden		
Versuchsgruppe	N	*M*	*SD*	N	*M*	*SD*	N	*M*	*SD*
Metakognitiver Prompt	108	2.36	0.79	108	2.37	0.76	101	2.39	0.75
Fehlererklärung	94	2.32	0.75	92	2.32	0.75	90	2.30	0.74
Inferenzprompt	96	2.32	0.84	95	2.32	0.84	94	2.32	0.85
Knowledge of Result	95	2.33	0.78	93	2.34	0.77	89	2.35	0.78
Kein Feedback	92	2.27	0.74	89	2.28	0.74	88	2.27	0.74
	$N_{gesamt} = 485$			$N_{gesamt} = 472$			$N_{gesamt} = 462$		
	F (4,480) = 0.16			F (4,467) = 0.17			F (4,457) = 0.34		
	$p > .05$			$p > .05$			$p > .05$		
	$\eta^2 = .001$			$\eta^2 = .001$			$\eta^2 = .003$		

ANOVAs für Testangst, Komponente Aufgeregtheit (3 Items), nach dem Experiment erhoben

	Cut-off-Kriterien: wann ein Item als zu schnell beantwortet gilt								
	< 7 Sekunden			< 8 Sekunden			< 9 Sekunden		
Versuchsgruppe	N	*M*	*SD*	N	*M*	*SD*	N	*M*	*SD*
Metakognitiver Prompt	108	1.60	0.65	103	1.60	0.65	101	1.61	0.65
Fehlererklärung	94	1.69	0.73	92	1.69	0.74	90	1.66	0.71
Inferenzprompt	96	1.63	0.72	95	1.64	0.72	94	1.64	0.72
Knowledge of Result	95	1.84	0.82	93	1.86	0.82	89	1.84	0.79
Kein Feedback	92	1.69	0.71	89	1.69	0.72	88	1.68	0.71
	$N_{gesamt} = 485$			$N_{gesamt} = 472$			$N_{gesamt} = 462$		
	F (4,480) = 1.60			F (4,467) = 1.71			F (4,457) = 1.50		
	$p > .05$			$p > .05$			$p > .05$		
	$\eta^2 = .013$			$\eta^2 = .014$			$\eta^2 = .013$		

ANOVAs für Einschätzungen zur Nützlichkeit der Rückmeldungen (5 Items), nach dem Experiment erhoben

	Cut-off-Kriterien: wann ein Item als zu schnell beantwortet gilt								
	< 7 Sekunden			< 8 Sekunden			< 9 Sekunden		
Versuchsgruppe	N	*M*	*SD*	N	*M*	*SD*	N	*M*	*SD*
Metakognitiver Prompt	108	2.80	0.58	103	2.83	0.57	101	2.89	0.57
Fehlererklärung	94	2.80	0.64	92	2.82	0.63	90	2.83	0.62
Inferenzprompt	96	2.77	0.50	95	2.77	0.50	94	2.78	0.50
Knowledge of Result	95	2.85	0.64	93	2.86	0.63	89	2.90	0.60
Kein Feedback	92	2.66	0.46	89	2.67	0.46	88	2.67	0.46
	$N_{gesamt} = 485$			$N_{gesamt} = 472$			$N_{gesamt} = 462$		
	F (4, 480) = 1.47			F (4, 467) = 1.64			F (4, 457) = 2.22		
	$p > .05$			$p > .05$			$p > .05$		
	$\eta^2 = .012$			$\eta^2 = .014$			$\eta^2 = .019$		

Anhang C:

Ergebnisse zentraler Analysen unter Nutzung der alternativen Cut-off-Werte für „Durchklicker“, die für Fallausschlüsse herangezogen werden

Analysen für Experiment 2

ANOVAs für Leistung in Erstversuchen (Treatmentphase), N = 31 Items

	Cut-off-Kriterien: wann ein Item als zu schnell beantwortet gilt								
	< 7 Sekunden			< 8 Sekunden			< 9 Sekunden		
Versuchsgruppe	N	*M*	*SD*	N	*M*	*SD*	N	*M*	*SD*
Inferenzprompt	58	14.41	5.43	57	14.58	5.33	57	14.58	5.33
Knowledge of Result	56	15.52	4.91	56	15.52	4.91	55	15.67	4.82
Kein Feedback	53	15.85	5.02	51	15.96	5.07	51	15.96	5.07
Inferenzprompt-via-Testleiter	59	17.42	4.47	59	17.42	4.47	59	17.42	4.70
	$N_{gesamt} = 226$ $F(3, 222) = 3.67$ $p < .05$ $\eta^2 = .05$			$N_{gesamt} = 223$ $F(3, 219) = 3.33$ $p < .05$ $\eta^2 = .044$			$N_{gesamt} = 222$ $F(3, 218) = 3.29$ $p < .05$ $\eta^2 = .043$		

Anmerkungen. Analog zu den im Hauptteil der Arbeit berichteten Ergebnissen wurden für die Berechnungen zusätzlich die Fälle mit unvollständigen Datensätzen des Experiments ausgeschlossen.

ANOVAs für Leistung in Zweitantworten (Treatmentphase) als relative Lösungshäufigkeiten (relativiert an Anzahl zweiter Versuche)

	Cut-off-Kriterien: wann ein Item als zu schnell beantwortet gilt								
	< 7 Sekunden			< 8 Sekunden			< 9 Sekunden		
Versuchsgruppe	N	*M*	*SD*	N	*M*	*SD*	N	*M*	*SD*
Inferenzprompt	58	0.41	0.16	57	0.42	0.15	57	0.41	0.15
Knowledge of Result	56	0.41	0.15	56	0.41	0.15	55	0.41	0.15
Inferenzprompt-via-Testleiter	59	0.49	0.17	59	0.49	0.17	59	0.49	0.17
	$N_{gesamt} = 173$ $F(2, 170) = 5.99$ $p < .01$ $\eta^2 = .066$			$N_{gesamt} = 172$ $F(2, 169) = 5.74$ $p < .01$ $\eta^2 = .064$			$N_{gesamt} = 171$ $F(2, 168) = 5.48$ $p < .01$ $\eta^2 = .061$		

Anmerkungen. Analog zu den im Hauptteil der Arbeit berichteten Ergebnissen wurden für die Berechnungen zusätzlich die Fälle mit unvollständigen Datensätzen des Experiments ausgeschlossen.

ANOVAs für Leistung in Posttest, N = 14 Items

	Cut-off-Kriterien: wann ein Item als zu schnell beantwortet gilt								
	< 7 Sekunden			< 8 Sekunden			< 9 Sekunden		
Versuchsgruppe	N	*M*	*SD*	N	*M*	*SD*	N	*M*	*SD*
Inferenzprompt	48	7.06	3.39	46	7.17	3.41	41	7.61	3.31
Knowledge of Result	49	8.10	2.54	47	8.30	2.40	43	8.58	2.31
Kein Feedback	46	7.07	3.09	44	7.18	3.09	42	7.21	2.92
Inferenzprompt-via-Testleiter	56	8.80	2.67	55	8.85	2.67	54	8.91	2.67
	$N_{gesamt} = 199$			$N_{gesamt} = 192$			$N_{gesamt} = 180$		
	F (3, 195) = 4.34			F (3, 188) = 4.13			F (3, 176) = 3.71		
	$p < .01$			$p < .01$			$p < .05$		
	$\eta^2 = .06$			$\eta^2 = .062$			$\eta^2 = .059$		

Anmerkungen. Analog zu den im Hauptteil der Arbeit berichteten Analysen wurden für die hier berichteten alternativen Berechnungen zusätzlich die Fälle mit unvollständigen Datensätzen des Posttests und des Experiments sowie die Fälle, die aufgrund des geltenden Kriteriums als „Durchklicker“ im Experiment identifiziert wurden, ausgeschlossen.

ANOVAs für Anstrengungsmotivation (1 Item), nach dem Experiment erhoben

	Cut-off-Kriterien: wann ein Item als zu schnell beantwortet gilt								
	< 7 Sekunden			< 8 Sekunden			< 9 Sekunden		
Versuchsgruppe	N	*M*	*SD*	N	*M*	*SD*	N	*M*	*SD*
Inferenzprompt	57	7.51	1.87	56	7.55	1.86	56	7.55	1.86
Knowledge of Result	55	7.78	1.60	55	7.78	1.60	54	7.78	1.61
Kein Feedback	53	7.32	2.25	51	7.33	2.29	51	7.33	2.29
Inferenzprompt-via-Testleiter	59	7.75	1.90	59	7.75	1.90	59	7.75	1.90
	$N_{gesamt} = 224$			$N_{gesamt} = 221$			$N_{gesamt} = 220$		
	F (3, 220) = 0.70			F (3, 217) = 0.62			F (3, 216) = 0.60		
	$p > .05$			$p > .05$			$p > .05$		
	$\eta^2 = .009$			$\eta^2 = .008$			$\eta^2 = .008$		

Anmerkungen. Analog zu den im Hauptteil der Arbeit berichteten Analysen betrifft die ANOVAs zur Anstrengungsmotivation die Fallausschlüsse aufgrund unvollständiger Datensätze des Experiments sowie die aufgrund des geltenden Kriteriums (siehe Tabelle) als „Durchklicker“ im Experiment identifizierten Fälle.

ANOVAs für Einschätzungen zur Nützlichkeit der Rückmeldungen (5 Items), nach dem Experiment erhoben

	Cut-off-Kriterien: wann ein Item als zu schnell beantwortet gilt								
	< 7 Sekunden			< 8 Sekunden			< 9 Sekunden		
Versuchsgruppe	N	*M*	*SD*	N	*M*	*SD*	N	*M*	*SD*
Inferenzprompt	57	2.67	0.66	56	2.68	0.67	56	2.68	0.67
Knowledge of Result	55	2.71	0.70	55	2.71	0.70	54	2.72	0.71
Inferenzprompt-via-Testleiter	59	3.12	0.52	59	3.12	0.52	59	3.11	0.52
	$N_{gesamt} = 171$			$N_{gesamt} = 170$			$N_{gesamt} = 169$		
	F (2, 168) = 8.36			F (2, 167) = 8.11			F (2, 166) = 7.96		
	$p < .001$			$p < .001$			$p < .01$		
	$\eta^2 = .09$			$\eta^2 = .089$			$\eta^2 = .087$		

Anmerkungen. Analog zu den im Hauptteil der Arbeit berichteten Analysen betrifft die ANOVAs zur Nützlichkeit die Fallausschlüsse aufgrund unvollständiger Datensätze des Experiments sowie die aufgrund des geltenden Kriteriums (siehe Tabelle) als „Durchklicker“ im Experiment identifizierten Fälle.