Diagnostische Fähigkeiten von Lehrkräften bei der Einschätzung von Schülerleistungen und Aufgabenschwierigkeiten bei Lernmedien mit instruktionalen Bildern

Nele McElvany[1], Sascha Schroeder[1], Tobias Richter[2], Axinja Hachfeld[1], Jürgen Baumert[1], Wolfgang Schnotz[3], Holger Horz[4] & Mark Ullrich[3]

[1]Max-Planck-Institut für Bildungsforschung, Berlin

[2]Universität zu Köln

[3]Universität Koblenz-Landau

[4]Fachhochschule Nordwestschweiz Olten

angenommen zur Veröffentlichung in der *Zeitschrift für Pädagogische Psychologie*

Autorenhinweis:

Das Projekt BiTe („Entwicklung und Überprüfung von Kompetenzmodellen zur integrativen Verarbeitung von Texten und Bildern“) wird von der Deutschen Forschungsgemeinschaft im Rahmen des Schwerpunktprogramms „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (SPP 1293) gefördert.

Zusammenfassung

Diagnostische Fähigkeiten von Lehrkräften gelten als wichtige Voraussetzungen für die adäquate Vorbereitung, Durchführung und Nachbereitung des schulischen Unterrichts. Für den Unterricht sind wiederum in vielen Fächern Lernmaterialien grundlegend, die Texte mit instruktionalen Bildern enthalten. Vor diesem Hintergrund diente die vorliegende Studie der Untersuchung von zentralen Forschungsfragen zu Niveau und Zusammenhängen der diagnostischen Fähigkeiten von Lehrkräften im Bereich der Text-Bild-Integration, zu möglichen lehrer- bzw. materialseitigen Moderatorvariablen sowie zu Determinanten im diagnostischen Urteilsprozess. Es nahmen 116 Lehrkräfte mit 48 Klassen der Stufen 5 bis 8 unterschiedlicher Schulformen an der Studie teil. Zentrale Ergebnisse waren eine schwache bis moderate Güte der diagnostischen Lehrerurteile bei einer Tendenz zur Unterschätzung der Schülerleistungen, ein heterogenes Befundmuster bezüglich der Zusammenhänge mit fachdidaktischem Wissen und Berufserfahrung sowie eine Bedeutung von gesamtmaterialbezogenen zusätzlich zu aufgabenspezifischen Schwierigkeitseinschätzungen im Urteilsprozess. Die Befunde werden im Hinblick auf Implikationen für die Praxis und weiteren Forschungsbedarf diskutiert.

Abstract

Teachers' diagnostic skills are important prerequisites for the planning, delivery, and evaluation of lessons. Learning materials incorporating instructional pictures are the basis for lessons in many subjects. Against this background, the present study investigated the levels and correlations of teachers' diagnostic skills in the area of text-picture integration, potential teacher- or material-specific moderator variables, as well as determinants of the diagnostic process. Participants were 116 teachers with 48 grade 5 to 8 classes in different school types. Teachers' diagnostic skills were found to be weak to moderate, with teachers tending to underestimate their students' performance. Heterogeneous patterns of results emerged for correlations with pedagogical content knowledge and teaching experience, and overall difficulty estimates were found to be relevant in addition to task-specific judgments within the diagnostic process. Implications for practice and further research are discussed.

1. Einleitung

Diagnostische Fähigkeiten gelten mit Blick auf das Ziel einer optimalen Ausrichtung des Unterrichts auf die Voraussetzungen der Schüler als zentraler Aspekt der Lehrerkompetenz (vgl. Hoge & Coladarci, 1989; Schrader, 1989; Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2004; Spinath, 2005). Akkurate diagnostische Urteile ermöglichen es, unter Berücksichtigung der Lernvoraussetzungen der Schüler den jeweiligen Fachunterricht möglichst adäquat zu planen, durchzuführen und nachzubereiten (Rogalla & Vogt, 2008). Dabei basiert der Unterricht in den meisten Fächern auf schriftlichen Lernmaterialien, die Texte mit instruktionalen Bildern enthalten. Diagnostische Fähigkeiten scheinen daher in diesem Bereich für Lehrkräfte besonders relevant zu sein.

Instruktionale Bilder können sowohl realistische als auch logische Bilder mit unterschiedlichem Abstraktheitsgrad sein (beispielsweise Fotographien, Tabellen, graphische Abbildungen oder Diagramme). Erfolgreicher Wissenserwerb setzt in diesen Fällen das Extrahieren von Informationen aus zwei unterschiedlichen Quellen und die Integration dieser Informationen zu einer kohärenten Gesamtrepräsentation voraus (Mayer, 2002; Schnotz & Bannert, 2003). Entsprechend stehen die Lehrkräfte der verschiedenen Fächer bei der Beurteilung von Materialien im Bereich des integrativen Lesens von Texten mit instruktionalen Bildern vor der Herausforderung, die Schülerkompetenzen und mögliche Verständnisschwierigkeiten sowohl in Bezug auf das Lesen des Textes als auch des Bildes und zusätzlich die notwendige Integrationsleistung korrekt einschätzen zu müssen (vgl. Hosenfeld, Helmke & Schrader, 2002; Lehmann et al., 2000).

Bisher fehlen jedoch Erkenntnisse zu den diagnostischen Fähigkeiten im Bereich der Text-Bild-Integration bei Lehrkräften. Dies betrifft sowohl die Frage, wie gut Lehrkräfte die entsprechenden Materialien und Schülerleistungen einschätzen können, als auch wie unterschiedliche Komponenten diagnostischer Fähigkeiten untereinander zusammenhängen.

Ungeklärt ist bisher auch, welche lehrer-, schüler- oder materialseitigen Merkmale die Urteile zu Texten und Bildern und deren Akkuratheit im diagnostischen Urteilsprozess moderieren.

1.1 Diagnostische Fähigkeiten von Lehrkräften

Definition und Komponenten

Vor dem Hintergrund der schulischen Rahmenbedingungen und Anforderungen des Unterrichtalltags treffen die Lehrkräfte im Laufe eines Schuljahres diagnostische Urteile zu verschiedenen Zwecken und mit unterschiedlichem Formalisierungsgrad (vgl. auch embedded assessment u.a. White & Gunstone, 1992). Die diagnostische Kompetenz von Lehrkräften bezeichnet dabei zunächst die Fähigkeit, Merkmale von Personen korrekt einzuschätzen, also die Urteilsgenauigkeit (Schrader, 2006; vgl. Helmke, Hosenfeld & Schrader, 2004). In einem erweiterten Verständnis wird neben der personenbezogenen Urteilsfähigkeit auch die Einschätzung von Aufgaben als Teil der diagnostischen Kompetenz von Lehrkräften berücksichtigt (Südkamp, Möller & Pohlmann, 2008; vgl. fachdidaktisches Wissen bei Baumert & Kunter, 2006).

Spinath (2005) wies vor dem Hintergrund geringer Interkorrelationen unterschiedlicher diagnostischer Komponenten nachdrücklich darauf hin, dass die diagnostische Kompetenz nicht als eindimensionales Persönlichkeitskonstrukt zu verstehen ist. Stattdessen sind unterschiedliche Teilkomponenten im Sinne von verschiedenen diagnostischen Fähigkeiten anzunehmen. Im deutschsprachigen Raum wurden diese vor allem anhand von den drei Urteilskomponenten Niveaukomponente, Differenzierungskomponente und Vergleichs- oder Rangkomponente (auch diagnostische Sensitivität) untersucht (Hosenfeld et al., 2002; Schrader, 1989; Spinath, 2005). Die Niveaukomponente bildet dabei die Urteilstendenz (Differenz zwischen geschätztem und empirischem Wert) im Hinblick auf eine Über- oder Unterschätzung im Sinne eines bias ab. Zusätzlich kann auch über die mittlere absolute Abweichung zwischen geschätzten und empirischen Werten der absolute

Urteilsfehler bestimmt werden (konzeptuell die Effizienz der Schätzung; vgl. Abweichungsmaß u.a. bei Südkamp et al., 2008). In Abhängigkeit von den jeweiligen Forschungsfragen werden unterschiedliche Urteilskomponenten für die Analysen herangezogen.

Bedeutung

Die diagnostischen Fähigkeiten werden, neben Klassenführungs-, didaktischer und fachwissenschaftlicher Kompetenz, als ein zentraler Aspekt der Lehrerexpertise beschrieben (z. B. Baumert & Kunter, 2006; vgl. auch Anders, Kunter, Brunner, Krauss, & Baumert, im Druck). Sie werden theoretisch als wesentliche Voraussetzung dafür angesehen, dass im Sinne der adaptiven Lehrkompetenz der jeweilige Fachunterricht vor dem Hintergrund der Lernvoraussetzungen der Schüler so geplant und durchgeführt werden kann, dass für möglichst viele Schüler bestmögliche Bedingungen für das Erreichen der Lernziele bestehen (Rogalla & Vogt, 2008). Die grundlegende theoretische Hintergrundannahme ist, dass ein diagnostischer Prozess stattfindet, als dessen Ergebnis dem Lehrer diagnostische Einschätzungen zu einer bestimmten Klasse bzw. bestimmten Schülern sowie zu dem Lernmaterial vorliegen. Die getroffene Einschätzung der Schülerleistungen in der zu unterrichtenden Klasse sollte handlungsleitend insbesondere auch die Auswahl und ggf. Modifikation von kognitiv herausfordernden und adäquat schwierigen Unterrichtsmaterialien (Texte, Bilder, Aufgaben) aus Schulbüchern, Arbeitsblattsammlungen etc. in der Unterrichtsplanungsphase sowie den Umgang mit den Materialien im Unterricht beeinflussen.

Bisher liegen nicht nur für den Bereich der Text-Bild-Integration kaum empirische Erkenntnisse zum Zusammenhang zwischen diagnostischen Fähigkeiten und unterrichtsbezogenem Handeln der Lehrkräfte vor. Aktuell berichteten Anders et al. (im Druck) für den Bereich der Mathematik, dass der aufgabenbezogene Urteilsfehler der Lehrkräfte mit dem kognitiven Aktivierungspotenzial von Klassenarbeitsaufgaben korrelierte

und der Urteilsfehler sowie die diagnostische Sensitivität zumindest geringe Effekte auf die Mathematikleistung in Klasse 10 bei Kontrolle der Leistung in Klasse 9 hatten (siehe auch die umfangreiche Studie von Helmke & Schrader, 1987).

Akkuratheit der diagnostischen Urteile

Studien, die Korrelationen zwischen Lehrerurteil und Schülerleistung verschiedener Leistungsbereiche ermittelten, berichteten durchschnittlich relativ genaue Einschätzungen der Lehrkräfte. In einer Metaanalyse von Hoge und Coladarci (1989) über 16 Studien lag die mittlere Korrelation bei $\bar{r}$ = .66 bei der Einschätzung verschiedener Schülerleistungen durch Lehrkräfte. Auch im Bereich Lesen wurden insgesamt akzeptable gemittelte Urteils-Kriteriums-Korrelationen gefunden (Bates & Nettelbeck, 2001: r = .62; Demaray & Elliott, 1998: r = .82). Hingegen zeigten Studien, die mittlere Differenzwerte zwischen Lehrerurteil und Leseleistungswerten bestimmten, häufig eine Überschätzung der Leseleistung durch die Lehrkräfte (Begeny, Eckert, Montarello & Storie, 2008; Hamilton & Shinn, 2003; vgl. aber Feinberg & Shapiro, 2003 für eine Unterschätzung). Für den Bereich der Text-Bild-Integration liegen bisher noch keine empirischen Erkenntnisse zu den entsprechenden diagnostischen Fähigkeiten der Lehrkräfte vor. Zudem zeichnete sich in Studien zur Einschätzung von allgemeinen Leseleistungen eine breite Streuung zwischen den Lehrkräften in ihrer Urteilsgenauigkeit ab, die auf die Notwendigkeit der Untersuchung von Determinanten diagnostischer Fähigkeiten verweist.

Moderatoren

Die Güte von Lehrerurteilen in einem bestimmten Inhaltsbereich kann neben unterschiedlichen Aspekten der Schülergruppe (z. B. Leistungsheterogenität) sowie einzelner Schüler auch durch Merkmale der Lehrkraft oder des zu beurteilenden Materials beeinflusst sein. Als wichtiger Aspekt der professionellen Handlungskompetenz von Lehrkräften wird

neben Einstellungen, Motivation und Selbstregulation das Wissen der Lehrkräfte angesehen (Baumert & Kunter, 2006). Dieses Wissen kann in Anlehnung an Shulman (1986) in Fachwissen, fachdidaktisches Wissen und pädagogisches Wissen differenziert werden. So könnte insbesondere allgemeines fachdidaktisches Wissen, das als zentrale Komponente von Lehrerkompetenz und als eine der Determinanten von Unterrichtsqualität gilt, ein wichtiger Moderator der Güte von diagnostischen Urteilen sein (siehe auch van Ophuysen, 2006). Wissen über den Bereich der Text-Bild-Integration müsste dabei in den Fachdidaktiken der Fächer verankert sein, in denen Lernmaterialien mit Texten und integrierten instruktionalen Bildern verwendet werden. Während die Einschätzung von Textmaterial traditionell vor allem als Expertise der Deutschlehrkräfte gesehen wird, sind in der Sekundarstufe I insbesondere auch die Lernmaterialien der Fächer Erdkunde und Biologie auf Texten mit instruktionalen Bildern aufbauend.

Insbesondere in der Praxis ist die Meinung anzutreffen, mit größerer Erfahrung (= längerer Berufserfahrung) würden höhere Kompetenzen einhergehen. Als Voraussetzung für die Verfestigung professioneller Expertise in der schulischen Praxis gelten langfristige systematische und reflektierte Unterrichtserfahrungen in Kombination mit dem Lernen von Kollegen, Weiterbildungsangeboten und Feedback (Baumert & Kunter, 2006; Bromme, 1997; Brunner, Kunter, Krauss & Baumert, 2006). Reine Erfahrung – in diesem Fall die langjährige Unterrichtstätigkeit – in einem Bereich sollten hingegen nicht ausreichen, um spezifische Kompetenzen – wie beispielsweise im Gebiet der Text-Bild-Integration – zu erwerben (vgl. Ericsson & Charness, 1994; Scardamalia & Bereiter, 1991). Für den Bereich des integrativen Lesens von Texten mit instruktionalen Bildern, der selten expliziter Fokus in der pädagogischen Praxis ist, ist allerdings kaum davon auszugehen, dass diese Bedingungen im schulischen Alltag zutreffen.

In Bezug auf weitere Moderatoren machten unter anderem Eckert, Dunn, Codding, Begeny und Kleinmann (2006) darauf aufmerksam, dass die Urteilsgüte zu Leseleistungen

nicht nur vom Schülerleistungsniveau abhingen, sondern auch von der Schwierigkeit des einzuschätzenden Materials. Probleme zeigten sich vor allem bei der akkuraten Einschätzung von Schülern, die Materialien, die für die untersuchte Klassenstufe oder höher eingestuft waren, auf dem höchsten Kompetenzniveau lasen.

Neben der Frage nach Moderatoren der diagnostischen Urteilsgüte (Akkuratheit) ist die Frage nach Determinanten der getroffenen Urteile selber zu stellen. Insbesondere in Bezug auf die Einschätzung von Aufgabenschwierigkeiten ist bisher ungeklärt, ob sich Lehrkräfte in ihrem Urteilsprozess neben spezifischen Einschätzungen auch von einem übergreifenden Gesamteindruck leiten lassen – analog zu den Anker-Effekten, die für die Einschätzung von Personen gefunden wurden (vgl. auch Schrader & Helmke, 1987, zu anderen eher überdauernden Lehrermerkmalen als Einflussfaktoren). Van Ophuysen (2006) berichtete, dass insbesondere Personen mit weniger Expertise zu einer Bestätigungstendenz (confirmation bias) von vorläufigen, selbst gefällten Urteilen neigten und spätere widersprüchliche Information weniger berücksichtigten - eine Tendenz, die in der kognitiven Sozialpsychologie durch eine spezifisch verzerrte Suche, Wahrnehmung und Bewertung von Informationen erklärt wird. Krolak-Schwerdt und Rummer (2005) zeigten, dass Experten, wenn es darum ging sich einen „Eindruck“ zu verschaffen (statt einer genauen Leistungsprognose), prototypische Kategorien und weniger aufmerksamkeitsintensive Strategien der Merkmalsintegration aktivierten und die Eindrucksbildung mit einer Top-down-Strategie einherging (siehe z.B. auch Fiske, Neuberg, Beattie & Milberg, 1987). Bei den berichteten Studien standen Schülerbeurteilungen im Mittelpunkt der Untersuchung und bisher ist ungeklärt, inwiefern die berichteten Ergebnisse auch für die Einschätzung von Materialien gelten.

1.2 Texte mit instruktionalen Bildern als Unterrichtsmaterialien

Bedeutung, Prozesse und Lehrerkompetenz

Das Lernen im schulischen Kontext beruht in den meisten Fächern auf Texten, die instruktionale Bilder enthalten, so dass die Lehrkräfte in diesem Bereich im Unterrichtsalltag vielfach diagnostische Einschätzungen treffen müssen. Bilder haben mit zunehmender Klassenstufe eine weniger dekorative und stärker instruktionale Funktion (z. B. Diagramme oder schematische Darstellungen). Experimentelle Studien belegten den Nutzen von instruktionalen Bildern als eigenständige, inhaltlich ergänzende Informationsquellen beim Verstehen von Inhalten und Lernen aus Texten und Textzusammenfassungen (Mayer, 2001; vgl. Schnotz & Kulhavy, 1994; Willows & Houghton, 1987).

Beim integrierten Lesen von Texten und Bildern werden die komplexen Prozesse des verstehenden Lesens um weitere Anforderungen erweitert (Mandl & Levin, 1989; Schnotz, 2005). Lesende stehen bei Texten mit integrierten Bildern vor der Herausforderung, zum einen Informationen aus den beiden Informationsquellen – Text und Bild – zu entnehmen und zum anderen diese sinnvoll miteinander zu verknüpfen, um unter Einbezug ihres Vorwissens ein mentales Gesamtverständnis der Inhalte aufbauen zu können (Ainsworth, 2006; Mayer, 2002; Schnotz & Bannert, 2003). Dabei wird davon ausgegangen, dass die Enkodierung der verbalen und piktorialen Informationen über unterschiedliche Kanäle mit jeweils begrenzter Verarbeitungskapazität stattfindet (vgl. Paivio, 1986; vgl. auch Chandler & Sweller, 1991; Mayer, 2001, 2002). Die integrative Verarbeitung beinhaltet ein Oberflächenstruktur-Mapping zwischen Wort- und Bildelementen (z. B. über Farbcodes oder Hinweispfeile) sowie ein Tiefenstruktur-Mapping durch einfache oder komplexe Relationen von sinngebendenen Einheiten beider Informationsquellen (Schnotz & Bannert, 2003).

Die Entwicklung von entsprechenden Kompetenzen zur integrativen Verarbeitung von Text- und Bildinformationen dürfte analog zur Kompetenzentwicklung in anderen Bereichen maßgeblich vom Unterricht und damit den Kompetenzen der Lehrkräfte beeinflusst werden (vgl. Houghton & Willows, 1987; allgemein Darling-Hammond, 2000). Dies erfordert insbesondere diagnostische Fähigkeiten der Lehrkräfte in diesem Bereich als Grundlage für

eine möglichst optimale Auswahl an Materialien entsprechend dem Leistungspotenzial einer Klasse und eine adäquate Instruktion im Unterricht (vgl. Hoge & Coladarci, 1989; Schrader, 1989; Spinath, 2005). Das integrative Lesen von Texten mit instruktionalen Bildern ist jedoch bisher kaum explizites Thema in den Curricula der Lehreraus- und –weiterbildung (vgl. auch Kremling, 2008).

Herausforderung bei der Diagnostik und relevante Diagnosebereiche und -ebenen

Neben der Fähigkeit der Lehrkräfte, Schülerkompetenzen beim integrativen Text-Bild-Lesen richtig einzuschätzen, dürfte für die Unterrichtsvorbereitung und -gestaltung insbesondere auch die Fähigkeit, Texte, instruktionale Bilder und die zugehörigen Aufgaben im Hinblick auf ihre Anforderungen, Schwierigkeit und Angemessenheit für bestimmte Zielgruppen korrekt einzuschätzen, grundlegend sein (vgl. Hosenfeld et al., 2002; siehe die Einordnung dieser Aspekte im Bereich fachdidaktisches Wissen bei Baumert & Kunter, 2006). Lehrkräfte der verschiedenen Fächer stehen dabei vor der Herausforderung, sowohl den Text als auch das Bild und zusätzlich die notwendige Integrationsleistung auf Seiten der Schüler im Hinblick auf die Schwierigkeit und Anforderung korrekt einzuschätzen. Dies gilt sowohl für eine übergreifende Gesamteinschätzung der Materialien als auch für die spezifische Beurteilung einzelner Aufgaben für die Schüler zu den Text-Bild-Materialien. Auch müssen im Hinblick auf die Schüler verschiedene Voraussetzungen für die Text-Bild-Integrationsprozesse berücksichtigt werden.

In diesem Kontext sind unterschiedliche Urteilsbereiche und -ebenen relevant (siehe Tabelle 1): Urteilsbereiche umfassen einerseits die Beurteilung der Schülerkompetenzen und andererseits die fachdidaktische Beurteilung von Unterrichtsmaterialien im Hinblick auf Schwierigkeiten und Anforderungen. Urteilsebenen können einzelne Schüler bzw. Aufgaben sein, aber auch ganze Klassen bzw. Tests. Eigenen Schülern/Klassen bzw. spezifischen Aufgaben/Tests übergeordnet ist die Ebene der allgemeinen übergreifenden Einschätzungen,

beispielsweise der zu erwartenden Fähigkeit von Schülern oder der Angemessenheit eines Unterrichtsthemas für eine bestimmte Jahrgangsstufe als Bezugsrahmen für den eigenen schulischen Kontext. Für jede Ebene können in beiden Bereichen die unterschiedlichen Komponenten diagnostischer Fähigkeiten untersucht werden (vgl. Abschnitt 1.1).

-- Bitte Tabelle 1 in etwa hier einfügen --

2. Forschungsfragen

Vor dem dargestellten Hintergrund diente die vorliegende Studie der Untersuchung von vier zentralen Forschungsfragen:

(1) Wie gut sind die diagnostischen Fähigkeiten von Lehrkräften der Sekundarstufe I im Bereich der Text-Bild-Integration und welche Zusammenhänge gibt es zwischen unterschiedlichen diagnostischen Fähigkeiten?

In den Curricula der Lehrerausbildung fehlt bisher in der Regel ein spezifischer Fokus auf Materialien mit Text-Bild-Integration, und in Studien zur Einschätzung der Textlesekompetenz wurden überwiegend Überschätzungen von Schülerleistungen berichtet (vgl. Abschnitte 1.1 und 1.2). Vor diesem Hintergrund wurde nur eine moderate und im Vergleich zu den berichteten Studien aus anderen Bereichen geringere Akkuratheit der diagnostischen Urteile von den Lehrkräften (Hypothese 1a) sowie eine Urteilstendenz zur Überschätzung der Schülerleistungen (Hypothese 1b) erwartet. Aufgrund der in anderen Studien gefundenen Heterogenität verschiedener diagnostischer Maße (z. B. Spinath, 2005) wurden auch für den hier untersuchten Bereich von geringen Interkorrelationen zwischen verschiedenen diagnostischen Maßen ausgegangen (Hypothese 1c).

(2) Sind fachdidaktisches Wissen über den Bereich der Text-Bild-Integration und Berufserfahrung als zentrale lehrerseitige Merkmale Moderatoren diagnostischer Fähigkeiten? Das fachdidaktische Wissen sollte in positivem Zusammenhang mit den diagnostischen Fähigkeiten stehen (vgl. Abschnitt 1.2; Hypothese 2a). Dabei war vor allem zu erwarten, dass

Wissen, das sich auf Aufgaben im Bereich der Text-Bild-Integration bezieht, insbesondere mit diagnostischen Fähigkeiten korreliert, die auf die explizite Beurteilung der Schwierigkeit von Aufgaben fokussieren. Vor dem Hintergrund der Erkenntnisse zum Kompetenzerwerb (vgl. Abschnitt 1.2) ist hingegen nicht davon auszugehen, dass die Berufserfahrung im Sinne von längerer Berufspraxis in Zusammenhang mit den diagnostischen Fähigkeiten steht (Hypothese 2b).

(3) Ist das generelle Schwierigkeitsniveau des Materials (Text, Bild, Verbindung Text-Bild) als wesentliches materialseitiges Merkmal ein Moderator der diagnostischen Akkuratheit?

Zum Einsatz kamen in der Untersuchung zur diagnostischen Einschätzung von Schülerleistungen und Aufgabenschwierigkeiten bei einem Teil der Stichprobe leichtes sowie bei dem anderen Teil der Stichprobe mittelschweres Material, das jeweils aus einem Text mit einem Bild und sechs zugehörigen Aufgaben bestand. Es wurde erwartet, dass Unterschiede in der diagnostischen Akkuratheit bei den aufgaben- und gesamttestbezogenen Urteilen festzustellen sind (Hypothese 3).

(4) Hat die allgemeine Schwierigkeitseinschätzung des Gesamtmaterials (Text, Bild, Verbindung Text-Bild) zusätzlich zu den aufgabenspezifischen Schwierigkeitseinschätzungen Bedeutung im diagnostischen Prozess als Determinanten der Lehrerurteile über die Aufgabenschwierigkeiten?

Basierend auf den bisherigen Erkenntnissen über diagnostische Urteile (vgl. Abschnitt 1.1) wurde ein zusätzlicher Einfluss der allgemeinen Schwierigkeitseinschätzungen vermutet (Hypothese 4).

3. Methode

3.1 Stichprobe

Die Untersuchung fand im Rahmen des BiTe-Projektes (Universität Koblenz-Landau/Max-Planck-Institut für Bildungsforschung, Berlin) im Februar 2008 an 48 Hauptschulen, Realschulen und Gymnasien in Rheinland-Pfalz statt. Die Schulen und innerhalb der Schulen jeweils eine Klasse aus der 5., 6., 7. oder 8. Klassenstufe wurden zufällig gezogen. Auf der Lehrerebene waren jeweils die Biologie-, Erdkunde- und Deutschlehrkräfte der Klassen beteiligt, da Text-Bild-Materialien in den Fächern Erdkunde und Biologie eine große Rolle spielen, während im Deutschunterricht vor allem mit Texten gearbeitet wird. Es lagen Daten von 116 von maximal möglichen 132 Lehrkräften vor. Die teilnehmenden Lehrkräfte waren im Durchschnitt 44.1 Jahre alt (SD = 11.4) und 65.1 % von ihnen waren weiblich. Die Lehrkräfte verteilten sich in etwa gleichmäßig auf die Schulformen (Hauptschule: 29.3%; Realschule: 34.5%; Gymnasium: 36.2%).

3.2 Instrumente

Diagnostische Fähigkeiten

Zur Erfassung der verschiedenen Komponenten der diagnostischen Fähigkeiten wurde den Lehrkräften ein Fragebogen vorgelegt, der repräsentatives Beispielmaterial aus einem Fähigkeitstest zur Bildtext-Integration enthielt, das von allen teilnehmenden Schülern der Studie bearbeitet wurde. Das Beispielmaterial bestand aus einem halbseitigen, nicht-kontinuierlichen Text mit einem zugehörigen instruktionalen Bild, für dessen Bearbeitung die gemeinsame Verarbeitung sowohl von Text- als auch von Bild-Informationen notwendig war. Zu dem Material gehörten sechs Multiple-Choice-Aufgaben, die das Verständnis von unterschiedlichen Informationen in dem Material abprüften. Thematisch wurden fachunspezifische Beispiele ausgewählt, die sich mit der Herstellung von Schokolade bzw. mit Hausmüll beschäftigten. Als erstes schätzten die Lehrkräfte global das allgemeine Schwierigkeitsniveau des Materials differenziert nach a) Text, b) instruktionalem Bild und c) Verbindung der Informationen aus Text und Bild für Schüler der untersuchten Klassenstufe und Schulform ein. Hierzu stand ihnen eine sechsstufige Antwortskala von 1 = sehr leicht

über 2 = ziemlich leicht, 3 = eher leicht, 4 = eher schwierig und 5 = ziemlich schwierig bis 6 = sehr schwierig zur Verfügung. Dann wurden insgesamt sechs Maße diagnostischer Fähigkeiten erhoben:

In Bezug auf die diagnostische Sensitivität gegenüber Aufgabenmerkmalen wurden die Lehrkräfte zunächst gebeten, die sechs Aufgaben des Materials ihrer Schwierigkeit nach in eine Rangreihenfolge zu bringen. Die resultierende Rangreihe wurde dann mit den in der Schülertestung empirisch ermittelten Aufgabenschwierigkeiten verglichen und für jede Lehrperson ein individueller Rangkorrelationskoeffizient berechnet (Rangkomponente sensu Schrader & Helmke, 1987). Diese wurden anschließend mittels einer Fisher's Z-Transformation über alle Lehrkräfte gemittelt. Positive Werte sprechen dabei generell für eine gute Einschätzung des relativen Schwierigkeitsgrades einer Aufgabe, Werte um den Nullpunkt oder negative Werte für eine nicht-systematisch bzw. den realen Aufgabenschwierigkeiten sogar entgegen gesetzte Einschätzung.

Zur Erfassung des aufgabenbezogenen Urteilsfehlers und der aufgabenbezogenen Urteilstendenz (Niveaukomponente) wurden die Lehrkräfte gebeten, für jede der sechs Aufgaben die prozentuale Lösungshäufigkeit in ihrer Klasse anzugeben. Die geschätzte Lösungshäufigkeit wurde dann von der in der Schülertestung empirisch ermittelten Lösungshäufigkeit abgezogen (Urteilstendenz). Positive Werte sprechen für eine tendenzielle Überschätzung der Schülerleistungen, negative für eine Unterschätzung. Die mittleren absoluten Abweichungen zwischen geschätzter und empirischer Lösungshäufigkeit geben den absoluten Urteilsfehler wieder, der sowohl Über- als auch Unterschätzungen beinhaltet. Je stärker dieser Wert von Null abweicht, desto größer der Fehler. Die Reliabilität der sechs Einschätzungen pro Lehrkraft lag für die Urteilstendenz bei Cronbach's $\alpha = .86$ und für den Urteilsfehler bei Cronbach's $\alpha = .61$.

Bei der Ermittlung des gesamttestbezogenen Urteilsfehlers sowie der gesamttestbezogenen Urteilstendenz wurde analog zu den aufgabenbezogenen Maßen

vorgegangen: Hier schätzten die Lehrkräfte die Gesamttestleistung ihrer Klasse ein (Anzahl der gelösten Aufgaben von insgesamt 48 Aufgaben im Gesamttest). Diese wurde von der empirisch ermittelten Klassenleistung abgezogen (Urteilstendenz) bzw. die absoluten Abweichungen berechnet (Urteilsfehler).

Zur Überprüfung der diagnostischen Sensitivität in Bezug auf die Schüler wurden die Lehrkräfte gebeten, sieben zufällig aus ihrer Klasse ausgewählte Schüler hinsichtlich ihrer Gesamttestleistung in eine Rangreihe zu bringen. Die Übereinstimmung mit der empirisch ermittelten Leistungsreihenfolge wurde wiederum mittels Rangkorrelation für jede Lehrkraft einzeln quantifiziert (Rangkomponente) und über eine Fisher's Z-Transformation gemittelt.

Wissen im Bereich Text-Bild-Integration – Multiple-Choice-Test

Der Multiple-Choice-Test zur Erfassung des Wissens der Lehrkräfte über den Bereich der Text-Bild-Integration umfasste 19 Aufgaben zu instruktionalen Bildern sowie diesbezüglichen Lese- und Instruktionsprozessen mit jeweils zwei oder vier Antwortalternativen (für Beispielaufgaben siehe Abbildung 1; vgl. auch McElvany et al., in Vorb.). Richtige Antworten wurden mit einem Punkt, falsche Antworten mit null Punkten gewertet. Hatte eine Person mehr als 50% fehlende Werte wurde sie von der Auswertung ausgeschlossen. Konsekutive fehlende Werte am Ende wurden als „not administered", andere fehlende Werte als falsch gewertet. Der Test hatte eine Reliabilität von Cronbachs $\alpha = .66$, die die Heterogenität des Wissenskonstrukts widerspiegelt.

-- Bitte Abbildung 1 in etwa hier einfügen --

Wissen über schwierigkeitsgenerierende Aufgabenmerkmale – Offene Antwort

Neben dem Multiple Choice-Test wurde das Wissen der Lehrkräfte auch im offenen Format abgefragt. Die einleitende Frage zur Erfassung des Wissens über schwierigkeitsgenerierende Aufgabenmerkmale lautete: „Welche Aufgabenmerkmale machen

Aufgaben zu einem Text mit instruktionalem Bild schwierig?“ Die Lehrkräfte konnten maximal zehn Antworten geben. Richtige Antworten wurden mit einem Punkt gescored, falsche oder irrelevante Antworten mit null Punkten (M = 1.21; SD = 1.61; Min = 0, Max = 7 Punkte). Die Antworten der Lehrkräfte wurden unabhängig von zwei Personen bepunktet, wobei die Interraterreliabilität bei Cohen’s Kappa = .80 lag.

Berufserfahrung

Die Berufserfahrung wurde über die Frage „Wie viele Jahre insgesamt (einschließlich der Referendariatszeit/ Vorbereitungszeit) werden Sie am Ende des Schuljahres unterrichtet haben?“ erfasst. Die teilnehmenden Lehrkräfte hatten eine durchschnittliche Unterrichtserfahrung von 16.6 Jahren (SD = 11.4; Spannbreite 2-40 Jahre).

3.3 Ablauf

Die Lehrkräfte beantworteten den Fragebogen, der u.a. die demografischen Angaben und die offenen Wissensfragen enthielt, zu Hause. Das Testheft mit den Aufgaben zu den diagnostischen Fähigkeiten sowie dem Multiple-Choice-Wissenstest bearbeiteten sie unter Aufsicht eines geschulten Testleiters in der Schule. Die Lehrkräfte der Klassenstufen 5 und 6 sowie 7 und 8 erhielten jeweils das gleiche Text-Bild-Material. Dabei war das Material für die 5. und 6. Klassenstufe einfach (durchschnittliche Lösungshäufigkeit der sechs Aufgaben zur Text-Bild-Integration: 86%, SD = 10%). Das Material für die 7. und 8. Klassen war mittelschwer (durchschnittliche Lösungshäufigkeit der sechs Aufgaben: 49%, SD = 16%). Die Lehrkräfte, die die einfachen bzw. mittelschweren Materialien beurteilten, unterschieden sich nicht nach Berufserfahrung ($t_{(106)} = 0.27$, $p > .05$), Wissen im Bereich Text-Bild-Integration (MC-Test; $t_{(106)} = 0.69$, $p > .05$) oder Geschlecht ($\chi^2_{(1)} = 0.35$, $p > .05$).

4. Ergebnisse

4.1 Niveau der diagnostischen Fähigkeiten und Zusammenhänge der Komponenten

Im Hinblick auf die Akkuratheit der diagnostischen Urteile (Hypothese 1a) zeigte sich, dass die Lehrkräfte der Sekundarstufe I die Aufgaben, die eine Text-Bild-Integration erforderten, nur maximal moderat korrekt im Hinblick auf unterschiedliche Schwierigkeit einschätzen konnten. Die mittlere Rangkorrelation der empirischen und der von den Lehrkräften eingeschätzten Aufgabenschwierigkeitsrangfolge lag bei $\bar{r}$ = .50 (SD = .31; diagnostische Sensitivität Aufgaben). Bei der Einschätzung der prozentualen Lösungshäufigkeit von sechs spezifischen Aufgaben in ihrer eigenen Klasse, die im Durchschnitt von 67% der Schüler gelöst wurden (SD = 23%), wichen die Lehrkräfte im Durchschnitt mit 2% (SD = 21%) kaum vom empirischen Mittelwert der Lösungshäufigkeiten ab (aufgabenbezogene Urteilstendenz) (siehe Abbildung 2). Eine weiterführende Betrachtung der absoluten Abweichungen (aufgabenbezogener Urteilsfehler) machte jedoch deutlich, dass die Abweichungen der prozentualen Lösungshäufigkeiten von den empirischen Mittelwerten nach oben und unten durchschnittlich bei 17% (SD = 13%) lagen.

-- Bitte Abbildung 2 in etwa hier einfügen --

Auch die Einschätzung der Gesamtzahl der von ihrer Klasse gelösten Aufgaben gelang insgesamt nur wenig präzise, wobei die Schüler durchschnittlich 33 der 48 Aufgaben lösten (SD = 6): Zwar unterschätzten die Lehrkräfte im Durchschnitt ihre Klassen bei insgesamt 48 Aufgaben nur um 3 Aufgaben (SD = 9; gesamttestbezogene Urteilstendenz). Die Betrachtung der absoluten Abweichungen vom empirischen Mittelwert nach oben und unten ergaben aber eine Fehleinschätzung von durchschnittlich 7 Aufgaben (SD = 6; gesamttestbezogener Urteilsfehler). Die in Hypothese 1b formulierte Annahme einer Überschätzung der Schülerleistungen kann demnach weder für die aufgaben- noch für die gesamttestbezogene Urteilstendenz bestätigt werden, für die eher eine Tendenz zur Unterschätzung der Schülerleistungen deutlich wurde. Die vergleichende diagnostische Beurteilung einzelner Schüler (diagnostische Sensitivität Schüler) war insgesamt nicht akkurat. Die mittlere

Rangkorrelation der empirischen und der von den Lehrkräften eingeschätzten Schülerleistungsrangfolge lag bei $\bar{r}$ = .34 (SD = .49).

Die Rangkorrelationen zwischen den verschiedenen diagnostischen Fähigkeiten (Hypothese 1c) waren in Abhängigkeit von ihrer konzeptuellen Nähe unterschiedlich hoch, wobei im Folgenden nur Korrelationen von mindestens mittlerer Höhe (r > .30) interpretiert werden (siehe Tabelle 2). Während es beispielsweise keinen statistisch signifikanten Zusammenhang zwischen der diagnostischen Sensitivität bei der Einschätzung der Aufgabenschwierigkeits- und bei der Schülerleistungsrangfolge gab, zeigte sich ein plausibler hoher Zusammenhang zwischen der Urteilstendenz bei einzelnen Aufgaben und der Urteilstendenz beim Gesamttest. Der absolute Urteilsfehler beruhte bei dem Gesamttest auf einer Unterschätzungstendenz sowohl bei der Einschätzung des Gesamttests als auch bei der Beurteilung der Schwierigkeit einzelner Aufgaben (vgl. negative Zusammenhänge in Tabelle 3). Interessant war in diesem Zusammenhang auch das Ergebnis, dass die Urteilstendenzen der Unterschätzung besonders dann höher ausgeprägt waren, wenn die diagnostische Sensitivität der Einschätzung einzelner Aufgaben vergleichsweise hoch war.

Unsere Daten unterstützten damit die theoretische Annahme multipler diagnostischer Fähigkeiten anstelle eines homogenen Konstrukts „diagnostische Kompetenz“ (Hypothese 1c). Insgesamt zeigten sich entsprechend der Hypothese 1a eher schwache bis moderate diagnostische Fähigkeiten der Lehrkräfte in der Sekundarstufe I bei einer Tendenz zur Unterschätzung der Kompetenzen der eigenen Klasse. Vermutet worden war im Vorfeld hingegen eine Überschätzung der Schülerleistungen (Hypothese 1b). Gleichzeitig wurde eine erhebliche Varianz in den diagnostischen Fähigkeiten zwischen den Lehrkräften festgestellt, die die Relevanz der Frage nach den Determinanten der diagnostischen Kompetenzen unterstreicht.

-- Bitte Tabelle 2 in etwa hier einfügen --

4.2 Wissen und Berufserfahrung als Moderatoren diagnostischer Fähigkeiten?

Hinsichtlich möglicher Moderatoren der diagnostischen Fähigkeiten der Lehrkräfte wurden zunächst die bivariaten Rangkorrelationen mit dem fachdidaktischem Wissen über den Bereich der Text-Bild-Integration untersucht. Dabei zeigte sich, dass die Zusammenhänge gering ausfielen. Die diagnostische Sensitivität bei der vergleichenden Einschätzung von Aufgaben bezüglich ihrer Schwierigkeit stand in statistisch signifikantem Zusammenhang mit dem Multiple-Choice-Test, der grundsätzliche Wissensaspekte zu instruktionalen Bildern, Leseprozessen und Anforderungen sowie Instruktionsprozessen umfasste (siehe Tabelle 2). Das Wissen über schwierigkeitsgenerierende Aufgabenmerkmale stand in statistisch signifikantem Zusammenhang mit der Urteilstendenz, die Kompetenzen der Schüler der eigenen Klasse beim Lösen bestimmter Aufgaben zu unterschätzen.

Die Analyse des Zusammenhangs von Berufserfahrung mit diagnostischen Fähigkeiten ergab gegenläufige Zusammenhangsmuster für die beiden Maße der diagnostischen Sensitivität (siehe Tabelle 2): Der statistisch signifikante Zusammenhang der Berufsdauer mit der diagnostischen Sensitivität bei der vergleichenden Einschätzung von individuellen Schülerleistungen war positiv. Lehrkräfte mit mehr Berufserfahrung konnten Schüler demnach besser in eine leistungsbezogene Rangreihe bringen. Lehrkräften, die länger in ihrem Beruf tätig waren, gelang es jedoch statistisch signifikant weniger gut, die Schwierigkeit der einzelnen Text-Bild-Aufgaben vergleichend einzuschätzen.

Insgesamt deutete das Ergebnismuster damit darauf hin, dass fachdidaktisches Wissen und Berufserfahrung lediglich in schwachem und nur teilweise zufallskritisch abzusicherndem Zusammenhang mit den diagnostischen Fähigkeiten im Bereich der Text-Bild-Integration standen. Während dieses Ergebnis für die Berufserfahrung in Einklang mit der im Vorfeld formulierten Hypothese (2b) steht, war für das fachdidaktische Wissen von einem positiven Einfluss ausgegangen worden (Hypothese 2a).

4.3 Schwierigkeit der Gesamtaufgabe als Moderator diagnostischer Akkuratheit?

Zunächst wurde ein Manipulation check durchgeführt, um zu überprüfen, ob die Lehrkräfte tatsächlich das leichtere Text-Bild-Material auch als leichter eingestuft haben als das schwerere Text-Bild-Material (siehe Tabelle 3). Bei der ANOVA waren die drei einzelnen Einschätzungen zum Schwierigkeitsniveau des Textes, des instruktionalen Bildes sowie des Verbindens der Informationen aus Text und Bild die abhängigen Variablen (AV) und die beiden Gruppen (leichtes bzw. mittelschweres Text-Bild-Material) die unabhängige Variable (UV). Das Ergebnis bestätigte, dass die Lehrkräfte die Materialien in der erwarteten Richtung statistisch signifikant unterschiedlich einstuften. Die Unterschiede waren mit $d = 0.47$ für die Texte, $d = 1.05$ für die Bilder und $d = 0.58$ für das Verbinden der Informationen von mittlerer bis großer Stärke. Damit wurde deutlich, dass insbesondere die Bilder als deutlich unterschiedlich schwer wahrgenommen wurden.

-- Bitte Tabelle 3 in etwa hier einfügen --

Die Überprüfung möglicher systematischer Unterschiede in der diagnostischen Akkuratheit in Abhängigkeit von der Schwierigkeit des Text-Bild-Materials (Hypothese 3) erfolgte mit einer ANOVA. Dabei bildeten die diagnostischen Fähigkeiten die AV und die beiden Text-Bild-Materialien die UV (siehe Tabelle 3). Bei den beiden Maßen der diagnostischen Sensitivität wurden die Z-transformierten Werte verwendet. Die Ergebnisse wiesen auf Unterschiede bei den Urteilstendenzen bei den Einschätzungen von einzelnen Aufgaben sowie von dem Gesamttest hin, wobei der Unterschied für den Gesamttest allerdings nicht mehr signifikant bleibt, wenn zusätzlich die Jahrgangsstufe der Schüler kontrolliert wird. Die Urteilstendenzen der Unterschätzung der prozentualen Aufgabenlösung waren bei den sehr leichten Materialien stärker ausgeprägt. Die vergleichende Einschätzung der Aufgabenschwierigkeit (diagnostische Sensitivität Aufgaben) gelang bei dem leichten Material statistisch signifikant besser als bei dem im Durchschnitt mittelschweren Material.

Keine statistisch signifikanten Unterschiede gab es hingegen bei den Urteilsfehlern und wie zu erwarten bei der diagnostischen Sensitivität bei der Schülereinschätzung. Damit verdeutlichten die Ergebnisse, dass das Schwierigkeitsniveau der zu berücksichtigenden Unterrichtsmaterialien für einige, aber nicht alle Aspekte diagnostischer Fähigkeiten als relevanter Moderator zu berücksichtigen ist und unterstützten damit teilweise Hypothese 3.

4.4 Diagnostischer Prozess: Bedeutung allgemeiner vs. aufgabenspezifischer Schwierigkeitseinschätzung

Die vierte Forschungsfrage beschäftigte sich mit der Frage, ob die allgemeine Schwierigkeitseinschätzung der Materialien (Text, Bild, Verbindung) für die diagnostischen Urteile zu der prozentualen Lösungshäufigkeit der sechs einzelnen Aufgaben in der eigenen Klasse zusätzlich zu den sechs aufgabenspezifischen Schwierigkeitseinschätzungen von prädiktiver Bedeutung ist. Hierzu wurde ein Prädiktionsmodell spezifiziert, in dem die allgemeine Schwierigkeitseinschätzung und die sechs aufgabenspezifischen Schwierigkeitseinschätzungen die Urteile zur Lösungshäufigkeit der sechs einzuschätzenden Aufgaben vorhersagten (siehe Abbildung 3). Korrelationen wurden zum einen zwischen den sechs aufgabenspezifischen Schwierigkeitseinschätzungen berücksichtigt. Zum anderen wurden Korrelationen zwischen den sechs Schwierigkeitseinschätzungen der einzelnen Aufgaben und der Gesamteinschätzung des Materials spezifiziert.

Die Modelle wurden in Mplus 5.1 (Muthén & Muthén, 1998-2008) gerechnet. Um alle zur Verfügung stehenden Informationen optimal zu nutzen und eine andernfalls mögliche Verzerrung der Ergebnisse zu vermeiden, wurde für die Analysen die FIML-Option gewählt (Full-Information Maximum-Likelihood; s.a. Lüdtke, Robitzsch, Trautwein & Köller, 2007). Zur Beurteilung der Modellgüte wurden der Comparative Fit Index (CFI) und der Root Mean Square Error of Approximation (RMSEA) herangezogen (Hu & Bentler, 1999).

Die Überprüfung der Vorhersagekraft der allgemeinen und aufgabenspezifischen Schwierigkeitseinschätzung (Hypothese 4) bestätigte zunächst erwartungsgemäß, dass die Urteile der Lehrkräfte über die Lösungshäufigkeiten einzelner Aufgaben in ihrer eigenen Klasse von ihrer Einschätzung der Schwierigkeit der einzelnen Aufgaben für die betreffende Jahrgangsstufe und Schulform statistisch signifikant vorhergesagt wurden (siehe Abbildung 3). Die allgemeine Einschätzung der Schwierigkeit des Materials (Text, Bild, Verbindung von Text-Bild) hatte aber darüber hinaus zusätzlich statistisch signifikante Erklärungskraft bei allen sechs Aufgaben. Die Modellanpassung war gut (χ^2 = 80.26; $\underline{df}$ = 54; $\underline{p} < .05$; CFI = .98; RMSEA = .07; $\underline{N}$ = 112). Ein Modellvergleich mit einem Modell, bei dem für jede der sechs Aufgaben die Koeffizienten der beiden Vorhersagepfade (allgemeine sowie aufgabenspezifische Einschätzung) gleichgesetzt wurden, passte weniger gut zu den Daten ($\Delta\chi^2$ = 23.92; $\underline{df}$ = 6; $\underline{p} < .001$). Der Einfluss der allgemeinen Schwierigkeitseinschätzung des Materials war demnach statistisch signifikant niedriger als der Einfluss der aufgabenspezifischen Schwierigkeitseinschätzung. Insgesamt kann damit aber in Einklang mit Hypothese 4 festgehalten werden, dass im diagnostischen Urteilsprozess bei den Lehrkräften über ihre aufgabenspezifischen Einschätzungen hinaus auch zusätzlich ihre allgemeinen Einschätzungen des Materials bedeutsam für die Urteile sind.

-- Abbildung 3 bitte in etwa hier einfügen --

5. Diskussion

Zusammenfassung und Interpretation

Die Studie erbrachte vier Hauptbefunde im Hinblick auf die diagnostischen Fähigkeiten von Lehrkräften bei Unterrichtsmaterialien mit Texten und instruktionalen Bildern: (1) Die Lehrkräfte unterschätzten tendenziell die Schülerkompetenzen und konnten die Materialien insgesamt nicht zufriedenstellend akkurat einschätzen. Die Akkuratheit lag dabei für einige Maße deutlich unter dem in anderen Studien berichteten Niveau (vgl. z.B. Hoge & Coladarci,

1989). Dies ist für die Praxis als problematisch anzusehen, da Text-Bild-Materialien in den meisten Fächern Unterrichtsgrundlage sind und die diagnostischen Fähigkeiten der Lehrkräfte als zentrale Voraussetzung für die adäquate Unterrichtsplanung und Unterrichtsdurchführung gelten (Rogalla & Vogt, 2008; Schrader, 2006). Der Befund der tendenziellen Unterschätzung der Schülerleistung steht dabei im Gegensatz zu der Mehrzahl publizierter Studien aus anderen Bereichen (z.B. Hosenfeld et al., 2002). Bei einer tendenziellen Überschätzung der Schülerleistungen kann argumentiert werden kann, dass dies aufgrund von stärkerer Anregung und Herausforderung auch positive Auswirkungen mit Blick auf die kognitive Lernentwicklung haben könnte. Bei einer Unterschätzung der Schülerinnen und Schüler ist zu befürchten, dass die vorhandenen Lernmaterialien mit Texten und instruktionalen Bildern nicht optimal einbezogen und genutzt werden, wenn die Schülerkompetenzen in diesem Bereich unterschätzt bzw. die Aufgaben selber in ihrer Schwierigkeit überschätzt werden. Dabei können fehlende diagnostische Fähigkeiten grundsätzlich in einer mangelnden Fähigkeit der Diagnostik der Schwierigkeit des Texts, der Schwierigkeit des Bildmaterials oder der Schwierigkeit der Integrationsleistung begründet liegen. In ergänzenden Studien mit modifiziertem Design müssen Rückschlüsse darauf gewonnen werden, welche Fähigkeitsfacetten für die hier gefundenen Urteilstendenzen bzw. diagnostische Sensitivität ursächlich sind.

Die geringen Interkorrelationen zwischen den verschiedenen Urteilskomponenten zeigten, dass nicht von einem einheitlichen Konstrukt „diagnostische Kompetenz“ ausgegangen werden kann (siehe bereits Spinath, 2005). Interessant ist in diesem Zusammenhang, dass die Unterschätzung der Schüler besonders dann höher ausgeprägt war, wenn die Lehrkräfte die Schwierigkeiten der Aufgaben besser differenzieren konnten – und analog, wenn sie vergleichsweise viel Wissen über schwierigkeitsgenerierende Aufgabenmerkmale hatten.

(2) Ein weiteres zentrales Ergebnis der Untersuchung waren die insgesamt nur schwachen Zusammenhänge zwischen fachdidaktischem Wissen bzw. Berufsdauer und diagnostischen

Fähigkeiten. Das heterogene Befundmuster eines positiven Zusammenhangs von längerer Berufsdauer mit der vergleichenden Einschätzung von Schülern im Gegensatz zu dem negativen Zusammenhang mit der Rangreihung von Aufgabenschwierigkeiten kann inhaltlich unterschiedlich interpretiert werden. Möglicherweise wurden Text-Bild-Materialien bei jüngeren Lehrkräften bereits stärker in der Ausbildung berücksichtigt oder ihre Bildrezeptionsgewohnheiten und –kompetenzen sind den Jugendlichen ähnlicher, so dass sie hier Vorteile bei der diagnostischen Einschätzung haben. Bei der zusammenfassenden Leistungsreihung der Schüler könnte hingegen die längere Berufspraxis und damit auch verbundene Erfahrungen mit Rückmeldungen zu getroffenen Schülereinschätzungen bei der Integration unterschiedlicher Informationen über die Schüler die Akkuratheit des Gesamturteils begünstigen.

(3) Die Ergebnisse zu der Frage, inwieweit das Schwierigkeitsniveau der einzuschätzenden Materialien ein Moderator der Akkuratheit ist, ergaben ein heterogenes Bild. Die Unterschiede in der diagnostischen Sensitivität bei der Aufgabeneinschätzung gingen deutlich in die Richtung, die weniger zu erwarten war: die Aufgaben, die leicht und von homogenem Schwierigkeitsniveau waren, wurden besser differenziert als die Aufgaben des mittelschweren Materials mit höherer Heterogenität bei den Aufgabenschwierigkeiten. Dass die Urteilstendenz bei dem leichten Material stärker war, kann auch unter methodischen Gesichtspunkten interpretiert werden, da eine starke Überschätzung der Schülerleistung hier weniger möglich ist. Andererseits berichteten Begeny et al. (2008), dass die Akkuratheit der Einschätzungen den Lehrkräften im Lesen besser gelang, wenn die Schüler eindeutig gut waren, als wenn sie mittlere bis schwache Leistungen zeigten. Möglicherweise ist die Akkuratheit der Einschätzung insgesamt eher durch Aspekte der Homogenität bzw. Heterogenität der Schülerleistungen innerhalb der Klasse und weniger der Aufgaben determinert (vgl. z.B. Karing, in diesem Heft; Rjosk, McElvany, Anders & Becker, in Vorb.; Weinert, Schrader & Helmke, 1990).

(4) Der vierte Hauptbefund ist die zusätzliche Vorhersagekraft, die der allgemeine Eindruck vom Schwierigkeitsniveau der Materialien über die aufgabenspezifischen Schwierigkeitseinschätzungen hinaus für die einzelnen getroffenen Urteile hat. Der Urteilsprozess beinhaltet demnach neben der Evaluation der aufgabenbezogenen Schwierigkeit auch zusätzlich eine bedeutsame Globalevaluationskomponente.

Einschränkungen, Implikationen und Ausblick

Nur ein Teil der insgesamt vorhandenen und relevanten diagnostischen Maße konnte in dieser Untersuchung berücksichtigt werden. Weitere Komponenten wie beispielsweise die diagnostische Einschätzung der Leistungen individueller Schüler bei einzelnen Aufgaben oder im Gesamttest gilt es in zukünftigen Studien zu untersuchen. Die überprüften Moderatorvariablen für die diagnostische Akkuratheit – Wissen, Berufserfahrung, Schwierigkeit des Materials – sind ebenso als erster Teilbereich zu sehen, dem weitere Analysen folgen müssen. Ein grundsätzliches Problem der diagnostischen Maße betrifft auch diese Studie: die in vielen Fällen nicht oder nur bedingt bestimmbare Reliabilität der Instrumente. Abschließend ist kritisch anzumerken, dass den Lehrkräften das Testformat sowohl auf Schüler- als auch auf Lehrerseite eher unbekannt war, was die Beurteilung erschwert haben könnte.

Diagnostische Fähigkeiten gelten als zentrale Voraussetzung für eine adäquate Unterrichtsgestaltung. Diesen angenommen Zusammenhang mit dem Unterricht gilt es in zukünftigen Studien zu untersuchen, die explizit thematisieren, inwieweit diagnostische Urteile handlungsleitend für die Lehrkräfte sind. Vor dem Hintergrund der hier berichteten schwachen Kompetenzen der Lehrkräfte in der Sekundarstufe I lässt sich als wichtige Implikation ableiten, dass der Erwerb diagnostischer Fähigkeiten im Bereich der Text-Bild-Integration einen stärkeren Fokus in Praxis wie Forschung benötigt. Hierzu gehört auch die Notwendigkeit, theoretisch fundierte, effektive Trainingsformate für Lehrkräfte in diesem

Bereich zu entwickeln. Dies könnte mittelfristig dann auch dazu beitragen, die berichteten Probleme der Schüler in Deutschland mit diesen Unterrichtsmaterialien zu verringern.

Literaturverzeichnis

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. Learning and Instruction, 16, 183-198.

Anders, Y., Kunter, M., Brunner, M., Krauss, S. & Baumert, J. (im Druck). Diagnostische Fähigkeiten von Mathematiklehrkräften und ihre Auswirkungen auf die Leistungen ihrer Schülerinnen und Schüler. Psychologie in Erziehung und Unterricht.

Bates, C. & Nettelbeck, T. (2001). Primary school teachers' judgements of reading achievement. Educational Psychology, 21, 177-187.

Baumert, J. & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. Zeitschrift für Erziehungswissenschaft, 9, 469-520.

Begeny, J. C., Eckert, T. L., Montarello, S. A. & Storie, M. S. (2008). Teachers' perceptions of students' reading abilities: An examination of the relationship between teachers' judgments and students' performance across a continuum of rating methods. School Psychology Quarterly, 23, 43-55.

Bromme, R. (1997). Kompetenzen, Funktionen und unterrichtliches Handeln des Lehrers. In F. E. Weinert (Hrsg.), Psychologie des Unterrichts und der Schule (S. 177-212). Göttingen: Hogrefe.

Brunner, M., Kunter, M., Krauss, S. & Baumert, J. (2006). Welche Zusammenhänge bestehen zwischen dem fachspezifischen Professionswissen von Mathematiklehrkräften und ihrer Ausbildung sowie beruflichen Fortbildung? Zeitschrift für Erziehungswissenschaft, 9, 521-544.

Chandler, P. & Sweller, J. (1991). Cognitive load theory and the format of instruction. Cognition and Instruction, 8, 293-332.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. Education Policy Analysis Archives, 8(1).

Demaray, M. K. & Elliott, S. N. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted performances. School Psychology Quarterly, 13, 8-24.

Eckert, T. L., Dunn, E. K., Codding, R. S., Begeny, J. C. & Kleinmann, A. E. (2006). Assessment of mathematics and reading performance: An examination of the correspondence between direct assessment of student performance and teacher report. Psychology in the Schools, 43, 247-265.

Ericsson, K. A. & Charness, N. (1994). Expert performance: Its structure and acquisition. American Psychologist, 49, 725-747.

Feinberg, A. B. & Shapiro, E. S. (2003). Accuracy of teacher judgements in predicting oral reading fluency. School Psychology Quarterly, 18, 52-65.

Fiske, S. T., Neuberg, S. L., Beattie, A. E. & Milberg, S. J. (1987). Category-based and attribute-based reactions to others: Some informational conditions of stereotyping and individuating processes. Journal of Experimental Social Psychology, 23, 399–427.

Hamilton, C. & Shinn, M. R. (2003). Characteristics of word callers: An investigation into the accuracy of teachers' judgments of reading comprehension and oral reading skills. School Psychology Review, 32, 228-240.

Helmke, A., Hosenfeld, I. & Schrader, F.-W. (2004). Vergleichsarbeiten als Instrument zur Verbesserung der Diagnosekompetenz von Lehrkräften. In R. Arnold & C. Griese (Hrsg.), Schulleitung und Schulentwicklung (S. 119-144). Hohengehren: Schneider-Verlag.

Helmke, A. & Schrader, F.-W. (1987). Interactional effects of instructional quality and teacher judgement accuracy on achievement. Teaching and Teacher Education, 3, 91-98.

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

Hosenfeld, I., Helmke, A. & Schrader, F.-W. (2002). Diagnostische Kompetenz: Unterrichts- und lernrelevante Schülermerkmale und deren Einschätzung durch Lehrkräfte in der Unterrichtsstudie SALVE. In M. Prenzel & J. Doll (Hrsg.), Bildungsqualität von Schule: Schulische und außerschulische Bedingungen mathematischer, naturwissenschaftlicher und überfachlicher Kompetenzen. Zeitschrift für Pädagogik, 45. Beiheft (S. 65-82). Weinheim: Beltz.

Houghton, H. A. & Willows, D. M. (Eds.). (1987). The psychology of illustration: Vol. 2. Instructional issues. Springer: New York.

Hu, L. T. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 5, 1-55.

Kremling, C. (2008). Erste Ergebnisse der Eingangserhebung des Forschungsprojektes " Bildliteralität und ästhetische Alphabetisierung". In G. Lieber (Hrsg.), Lehren und Lernen mit Bildern (S. 115-123). Schneider Verlag Hohengehren: Baltmannsweiler.

Krolak-Schwerdt, S. & Rummer, R. (2005). Der Einfluss von Expertise auf den Prozess der schulischen Leistungsbeurteilung. Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 37, 205-213.

Lehmann, R. H., Peek, R., Gänsfuß, R., Lutkat, S., Mücke, S. & Barth, I. (2000). Qualitätsuntersuchungen an Schulen zum Unterricht in Mathematik (QuaSUM). In Ministerium für Bildung, Jugend und Sport (Hrsg.), Schulforschung in Brandenburg, Heft 1. Teltow: Druckerei Grabow.

Lüdtke, O., Robitzsch, A., Trautwein, U. & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung: Probleme und Lösungen. Psychologische Rundschau, 58, 103-117.

Mandl, H. & Levin, J. R. (Hrsg.). (1989). Knowledge acquisition from text and pictures. Amsterdam: Elsevier.

Mayer, R. E. (2001). Multimedia learning. New York: Cambridge University Press.

Mayer, R. E. (2002). Using illustrations to promote constructivist learning from science text. In J. Otero, J. A. León & A. C. Graesser (Eds.), The psychology of science text comprehension (pp. 333–355). Mahwah, NJ: Lawrence Erlbaum.

McElvany, N., Schroeder, S., Richter, T., Hachfeld, A., Baumert, J., Schnotz, W., Horz, H. & Ullrich, M. (in Vorb.). Texte mit instruktionalen Bildern als Unterrichtsmaterial – Kompetenzen der Lehrkräfte.

Muthén, L. K. & Muthén, B. O. (1998-2008). Mplus user's guide. Los Angeles, LA: Muthén & Muthén.

Paivio, A. (1986). Mental representations: A dual coding approach. Oxford: Oxford University Press.

Rjosk, C., McElvany, N., Anders, Y. & Becker, M. (in Vorb.). Diagnostische Fähigkeiten von Lehrkräften bei der Einschätzung der basalen Lesefähigkeit ihrer Schülerinnen und Schüler.

Rogalla, M. & Vogt, F. (2008). Förderung adaptiver Lehrkompetenz: eine Interventionsstudie. Unterrichtswissenschaft. Zeitschrift für Lernforschung, 36, 17-36.

Scardamalia, M. & Bereiter, C. (1991). Literate expertise. In K. A. Ericsson & J. Smith (Eds.), Toward a general theory of expertise: Prospects and limits (pp. 172-194). Cambridge: Cambridge University Press.

Schnotz, W. (2005). An Integrated Model of Text and Picture Comprehension. In R. E. Mayer (Ed.), Cambridge Handbook of Multimedia Learning (pp. 49-69). Cambridge: Cambridge University Press.

Schnotz, W. & Bannert, M. (2003). Construction and Interference in Learning from Multiple Representation. Learning and Instruction, 13, 141-156.

Schnotz, W. & Kulhavy, R. W. (Eds.). (1994). Comprehension of graphics. Volume in the series Advances in Psychology. Amsterdam: Elsevier.

Schrader, F.-W. (1989). Diagnostische Kompetenz von Lehrern und ihre Bedeutung für die Gestaltung und Effektivität des Unterrichts. Frankfurt am Main: Lang.

Schrader, F.-W. (2006). Diagnostische Kompetenz von Eltern und Lehrern. In D. H. Rost (Hrsg.), Handwörterbuch Pädagogische Psychologie (3. überarb. u. erw. Aufl., S. 95-100). Weinheim: Beltz.

Schrader, F.-W. & Helmke, A. (1987). Diagnostische Kompetenz von Lehrern: Komponenten und Wirkungen. Empirische Pädagogik, 1, 27-52.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2004). Standards für die Lehrerbildung: Bildungswissenschaften. Beschluss der Kultusministerkonferenz vom 16.12.2004. Bonn.

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

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

Südkamp, A., Möller, J. & Pohlmann, B. (2008). Der Simulierte Klassenraum: Ein Instrument zur Untersuchung von diagnostischer Kompetenz. In E.-M. Lankes (Hrsg.), Pädagogische Professionalität als Gegenstand Empirischer Forschung (S. 87-97). Münster: Waxmann.

van Ophuysen, S. (2006). Vergleich diagnostischer Entscheidungen von Novizen und Experten am Beispiel der Schullaufbahnempfehlung. Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 38, 154-161.

Weinert, F. E. Schrader, F.-W. & Helmke, A. (1990). Educational expertise. Closing the gap between educational research and classroom practice. School Psychology International, 11, 163-180.

White, R. & Gunstone, R. (1992). Probing understanding. London: The Falmer Press.

Willows, D. M. & Houghton, H. A. (Eds.). (1987). The psychology of illustration: Vol. 1. Basic research. Springer: New York.

Tabelle 1

Systematisierung relevanter Urteilsbereiche und Urteilsebenen

		Urteilsbereich	
		Schüler z. B. Kompetenzen	Unterrichtsmaterial z. B. Anforderungen, Schwierigkeit
Urteilsebene	Individuelle Schüler		Einzelne Aufgaben
	Klasse		Gesamttest
	Übergreifend z. B. Klassenstufe/Schulform		Übergreifend z. B. Themenbereiche

Tabelle 2

Rangkorrelationen zwischen den diagnostischen Fähigkeiten sowie Zusammenhänge mit Wissen und Berufsdauer

		1	2	3	4	5	Wissen		Berufs-dauer
							Bilder, Prozesse, Instruktion	Aufgaben-merkmale	
1	Diagnostische Sensitivität: Aufgaben	1					.25*	.17^{+}	-.23*
2	Niveau-komponente: Urteilsfehler Aufgaben	-.04	1				-.07	.03	-.02
3	Niveau-komponente: Urteilstendenz Aufgaben	-.30**	-.12	1			-.06	-.20*	-.09
4	Niveau-komponente: Urteilsfehler Gesamttest	.20*	.29*	-.36***	1		-.01	.00	.13
5	Niveau-komponente: Urteilstendenz Gesamttest	-.23*	-.16*	.61***	-.54***	1	-.08	-.19^{+}	-.09
6	Diagnostische Sensitivität: Schüler	-.06	.23*	-.14	.14	-.16	-.07	.19^{+}	.21*

Anmerkungen. $^{+}$< .10, * p < .05, ** p < .01, *** p < .001; N = 91 - 112.

Tabelle 3

Einschätzungen der Lehrkräfte zur Schwierigkeit der Text-Bild-Materialien und diagnostische Fähigkeiten in Abhängigkeit von dem Schwierigkeitsniveau der Materialien

	M bzw. $\bar{r}$ (SD)		ANOVA
Einschätzung der Schwierigkeit	Text A (einfach) N = 54	Text B (mittelschwer) N = 57	
Text	3.04 (1.06)	3.53 (1.02)	$F(1,110) = 6.67$, $p < .05$, $eta^2 = .06$
Instruktionales Bild	2.87 (0.99)	3.98 (1.13)	$F(1,110) = 31.64$, $p < .001$, $eta^2 = .22$
Verbinden der Informationen aus Text und Bild	3.22 (1.16)	3.89 (1.15)	$F(1,109) = 9.45$, $p < .01$, $eta^2 = .08$
Diagnostische Fähigkeiten			
Diagnostische Sensitivität Aufgaben	.60 (.25)	.39 (.31)	$F(1,110) = 29.42$, $p < .001$, $eta^2 = .21$
Urteilsfehler Aufgaben	16.76 (14.07)	16.67 (11.86)	$F(1,108) = 0.001$, $p > .05$, $eta^2 = .00$
Urteilstendenz Aufgaben	-14.90 (16.06)	9.54 (18.19)	$F(1,108) = 55.47$, $p < .001$, $eta^2 = .34$
Urteilsfehler Gesamttest	7.39 (6.23)	6.73 (6.16)	$F(1,105) = 0.30$, $p > .05$, $eta^2 = .00$
Urteilstendenz Gesamttest	-5.73 (7.81)	-1.45 (9.05)	$F(1,105) = 6.79$, $p < .05$, $eta^2 = .06$
Diagnostische Sensitivität Schüler	.29 (.49)	.39 (.50)	$F(1,100) = 1.15$, $p > .05$, $eta^2 = .01$

Titel der Abbildung

Abbildung 1. Beispielaufgaben aus dem MC-Test zur Erfassung des Wissens der Lehrkräfte im Bereich der Text-Bild-Integration.

Abbildung 2. Aufgabenbezogene Urteilstendenzen der individuellen Lehrkräfte.

Abbildung 3. Vorhersage der diagnostischen Urteile aus den aufgabenspezifischen und gesamtmaterialbezogenen Schwierigkeitseinschätzungen.

Es gibt verschiedene Arten von Vergleichen, die in Aussagen vorkommen können. Welche Abbildungsart eignet sich <u>am besten</u>, wenn ein Lehrer seinen Schüler/innen die folgenden Vergleiche jeweils möglichst anschaulich darstellen möchte?

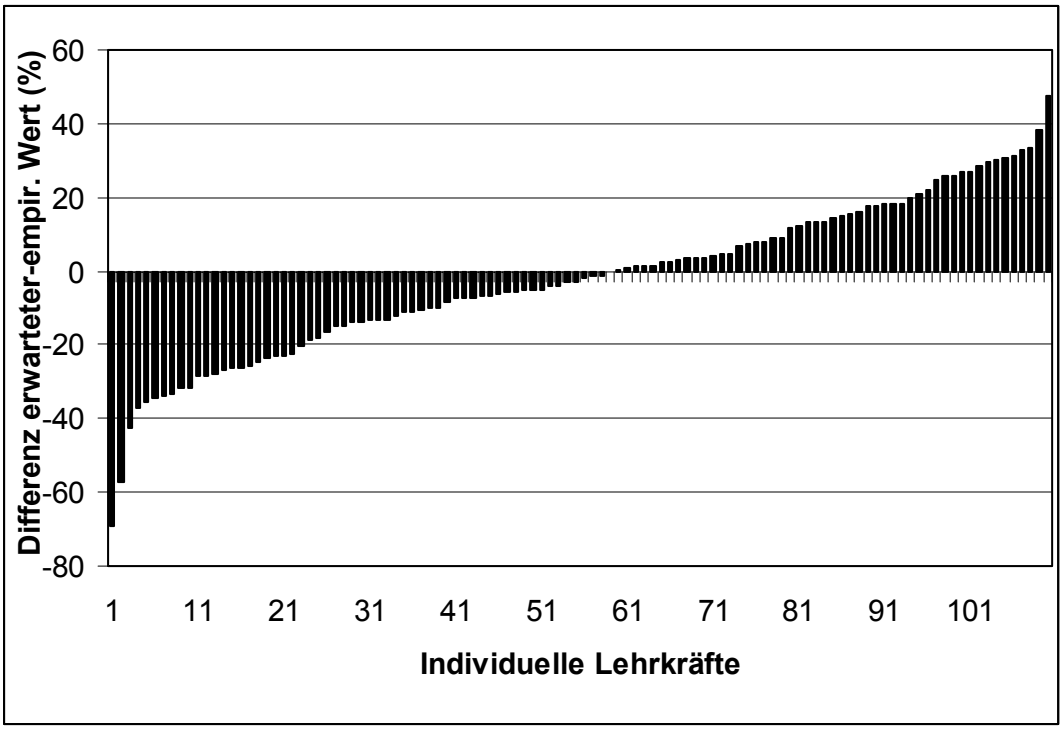
Darstellung…

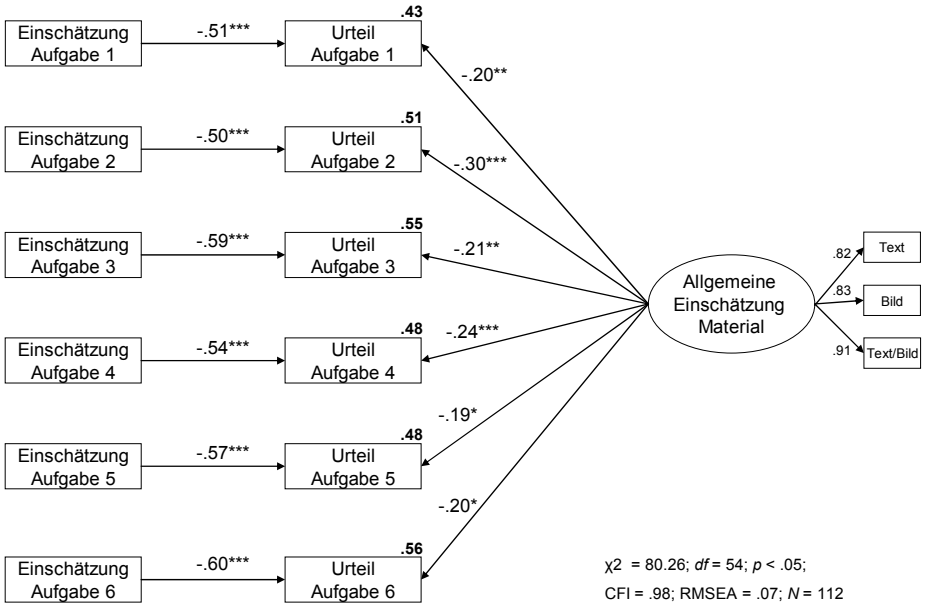
	Kreis-diagramm	Säulen-/ Balken-diagramm	Linien-diagramm	Punkte-diagramm
des Verlaufs einer Entwicklung	0	0	X	0
eines Teilmengenvergleichs.....	0	X	0	0

Welche der folgenden Aussagen zum Verarbeitungsaufwand bei der Abbildungs- und Textverarbeitung trifft zu?

	trifft <u>nicht</u> zu	trifft zu
Ein **Text**, der mit einer *komplexen* Abbildung illustriert ist, wird von guten Lesern intensiver verarbeitet als ein Text, der mit einer *einfachen* Abbildung illustriert ist.	0	X
Eine *einfache* **Abbildung**, die einen Text illustriert, wird intensiver verarbeitet als eine *komplexe* Abbildung, die einen Text illustriert.	X	0

<u>Anmerkung:</u> Die richtigen Antworten sind mit „X“ gekennzeichnet.


Differenz erwarteter-empir. Wert (%)
60
40
20
0
-20
-40
-60
-80
1
11
21
31
41
51
61
71
81
91
101
Individuelle Lehrkräfte



Anmerkung. Ohne Darstellung der Korrelationen zwischen den aufgabenspezifischen Schwierigkeitseinschätzungen (r = .55 - .87) sowie zwischen den Schwierigkeitseinschätzungen der einzelnen Aufgaben und der Gesamteinschätzung des Materials (r = .49 - .69). Fettgedruckte Werte beziffern R^2.