Julia Warwas / Tobias Kärner / Klaudia Golyszny

Diagnostische Sensibilität von Lehrpersonen im Berufsschulunterricht: Explorative Prozessanalysen mittels Continuous-State-Sampling

Kurzfassung: Lehrkräfte erbringen schülerbezogene Diagnoseleistungen nicht nur summativ-anlassbezogen wie etwa bei der Zeugniserstellung. Auch während des Unterrichts müssen sie das Verhalten der Schüler/innen kontinuierlich bewerten, um hieraus Schlüsse über Lernfortschritte zu ziehen und didaktische Feinjustierungen vornehmen zu können. Obwohl die Treffsicherheit solcher Urteile als unverzichtbare Voraussetzung der Unterrichtsqualität gilt, wurde sie bislang kaum prozessbegleitend in natürlichen Settings untersucht. Der vorliegende Beitrag setzt sich zunächst literaturgestützt mit den spezifischen Anforderungen pädagogischer Prozessdiagnosen und möglichen Bedingungsfaktoren intraindividuell schwankender Einschätzungsgenauigkeiten auseinander. Anschließend werden detaillierte Einzelfallanalysen von zwei Berufsschullehrkräften berichtet, deren klassenbezogene Urteile über das aktuelle Verstehen und die situative Langeweile im Verlauf von je neun Unterrichtsstunden engmaschig in zehnminütigen Intervallen erfasst und mit korrespondierenden Selbsteinschätzungen der Schüler/innen verglichen wurden. Die statistischen Auswertungen erhellen nicht nur, wie akkurat diese Lehrkräfte didaktisch relevante Schülermerkmale einzuschätzen vermögen, sondern auch, wie lehrerseitig wahrgenommene und videografierte Unterrichtsmerkmale die Urteilsgenauigkeit beeinflussen. Unter anderem zeigt sich, dass der leistungsnahe State-Parameter des aktuellen Verstehens grundsätzlich exakter eingeschätzt wird als derjenige des situativen Langeweileempfindens, und dass eine hohe Urteilsgenauigkeit für diesen Parameter vor allem im dialogischen Austausch mit den Schüler/innen, d.h. in Phasen des Lehrgesprächs, erreicht wird. Einer hypothesengenerierenden Logik entsprechend werden zuletzt Anregungen für künftige Untersuchungen gegeben.

Abstract: Teachers' tasks of assessing student characteristics are not confined to specific occasions, for example when issuing formal and summative certificates. Instead, teachers have to monitor and judge students' behavior continuously in each and every lesson in order to draw conclusions on their learning progress and on adequate didactical reactions. Although the accuracy of such judgements is deemed to be an indispensable requisite of teaching quality, it has seldom been examined empirically through in-process and in-situ measurements. This article first discusses the distinctive requirements of continuous pedagogical assessments and the conditional factors of intra-individual variations in judgement accuracy. Then, it presents findings from detailed case studies in which the judgements of two vocational teachers on students' situational understanding and boredom – measured in intervals of 10 minutes in the course of nine lessons each – were compared to corresponding self-assessments of their students. Statistical analyses do not only reveal how well these teachers manage to assess class-specific levels of understanding and boredom at each point in time, but also how their judgement accuracy is affected by both subjectively perceived and videotaped characteristics of ongoing instructional processes. Results suggest, among other things, that teachers judge varying states of students' understanding more accurately than varying states of boredom and that accuracy is particularly high in phases of dialogic exchange with the students, i.e. during guided classroom dialogue. In line with a hypothesis-generating approach, fruitful directions of follow-up studies are proposed.

1 Diagnostische Sensibilität als Voraussetzung guten Unterrichts und Gegenstand empirischer Forschung

Die von Lehrpersonen im Rahmen der *pädagogischen Diagnostik* zu erfüllenden Aufgaben sind vielfältiger Natur (im Überblick Ingenkamp & Lissmann 2008; Schrader 2013). Sie beinhalten zum einen regelmäßige Beurteilungen der Lernzielbeherrschung ihrer Schüler/innen, welche zum Zweck der *Ergebnisdiagnostik* überwiegend systematisch vollzogen und dokumentiert werden (z.B. Breuer 2010; Tenberg 2012). Sie umschließen zum anderen all jene Urteile, die sich im Dienste einer *Prozessdiagnostik* auf die gelingende Umsetzung geplanter Lehr-Lern-Prozesse in der laufenden Interaktion mit den Schüler/innen richten und daher permanent, jedoch mangels verbindlicher Richtlinien und Maßstäbe meist intuitiv-routinebasiert erfolgen (z.B. Beck et al. 2008). Diagnoseaufgaben sind folglich nicht nur punktuell und summativ anlässlich von *Selektions- und Allokationsentscheidungen* zu erbringen, wie sie z.B. für die Zeugniserstellung am Ende einer Qualifizierungsmaßnahme typisch sind. Vielmehr liefern sie in jeder Unterrichtseinheit wesentliche Informationen für (subjektiv) situationsangemessene und bedarfsgerechte Entscheidungen der *didaktischen Feinsteuerung*, bspw. für die Platzierung von Übungsphasen, die Formulierung von Impulsfragen oder die Konkretisierung von Rückmeldungen auf Arbeitsergebnisse. Im Idealfall ermöglicht eine ausgeprägte prozessdiagnostische Sensibilität eine hochgradig adaptive Unterrichtsgestaltung, die heterogenen Fähigkeiten, Bedürfnissen und Schwierigkeiten der Schüler/innen mit binnendifferenzierenden Lern- und Unterstützungsangeboten begegnet (van Buer & Zlatkin-Troitschanskaia 2009).

Obwohl die Treffsicherheit von Lehrendenurteilen als Kernvariable der Unterrichtsqualität gilt, welche die Effektivität stützender Maßnahmen wie etwa der Rückmeldehäufigkeit moderiert, wurde sie bislang kaum mithilfe *prozessbegleitender Erhebungen* in natürlichen oder simulierten Unterrichtssettings empirisch analysiert (van Ophuysen 2010; Helmke 2012; Behrmann & Souvignier 2013). In der gegenwärtigen Forschungslandschaft dominieren *status- und produktorientierte* Ansätze, die primär für die Einschätzung schulischer Leistungen reichhaltige Befunde in querschnittlich-komparativen Designs generiert haben (vgl. die Meta-Analyse von Südkamp, Kaiser & Möller 2012). Im Fokus stehen damit zeitpunktbezogene Urteile über potenzielle Performanzen der Schüler/innen in zentralen Prüfungsfächern (z.B. Lösung von Mathematikaufgaben), seltener auch über relativ zeitstabile kognitive oder emotional-motivationale Dispositionen (z.B. Intelligenz, fachliches Interesse oder Schulangst). Nur rudimentär erschlossen sind kontinuierliche Bewertungen unterrichtlicher Bedingungsvariationen sowie lernfortschrittsrelevanter, oszillierender ‚State-Parameter' der Schüler/innen, darunter aufgabenspezifische Unter-/Überforderung, situative Interessiertheit oder aktuelles Verstehen der soeben behandelten Unterrichtsinhalte (Ricken 2006)[1].

Zu den wenigen Ausnahmen zählt der Einsatz verschriftlichter oder gefilmter *Vignetten*, anhand derer pädagogische Experten und Novizen ausgewählte Aspekte

1 Der Vollständigkeit halber sei an dieser Stelle die pädagogische Lernprozessdiagnostik von der psychometrisch fundierten, *formativen Leistungsdiagnostik* abgegrenzt. Letztere betrachtet gerade nicht die Güte lehrerseitiger Einschätzungen von Unterrichts- bzw. Schülermerkmalen während des Lehr-Lern-Geschehens; vielmehr entlastet sie Lehrkräfte bei dieser Aufgabe, indem sie ihnen vermittels standardisierter Testverfahren in kurzen, regelmäßigen Zeitabständen reliable und valide Daten zur Leistungsentwicklung von Schüler/innen zur Verfügung stellt (vgl. hierzu Klauer 2014).

des dokumentierten Lehrerhandelns bewerten (Oser, Heinzer & Salzmann 2010; Seidel & Prenzel 2007), didaktische Pläne entwerfen und realisierte Vorgehensweisen begründen (Brühwiler 2014) oder aber domänenspezifische Schülerfehler und erfolgsträchtige Strategien der Fehlerbearbeitung identifizieren (Wuttke & Seifried 2013). Vereinzelt wurden zudem experimentelle Anordnungen im *computersimulierten Klassenraum* realisiert, um Referenzgruppen- oder Ankereffekte bei der Urteilsbildung über fiktive Schülerantworten kontrolliert zu überprüfen (Südkamp & Möller 2009; Dünnebier, Gräsel & Krolak-Schwerdt 2009).

Die hier berichtete explorative Studie macht sich das Verfahren des *Continuous-State-Samplings* zunutze, um in realen Interaktionsprozessen während des Berufsschulunterrichts einerseits die Güte prozessdiagnostischer Urteile und andererseits situative Einflüsse auf die Urteilsgüte zu analysieren. Gegenstände der Urteilsbildung sind dabei das *aktuelle Verstehen* der Lernenden als kognitiv akzentuiertes Merkmal sowie die *situative Langeweile* als emotional-motivationales Merkmal, die beide den Aufbau von Wissensstrukturen nachweislich begünstigen bzw. behindern können (z.B. Kögler, Bauer & Sembill 2011; Seifried 2004; Klieme & Rakoczky 2008).

2 Komponenten und Bedingungsfaktoren der Diagnosegüte

Per definitionem bezeichnet die diagnostische Kompetenz einer Lehrperson ihre Fähigkeit, Merkmale von Schüler/innen zutreffend einzuschätzen (Schrader 2010); sie kann sich aber ebenso in korrekten Bewertungen von Aufgabenmerkmalen niederschlagen (z.B. McElvany et al. 2009). Zur Quantifizierung des Ausmaßes, in dem solche Urteile den faktischen bzw. mit wissenschaftlichen Verfahren gemessenen Gegebenheiten entsprechen, dienen üblicherweise drei Kennwerte (Schrader & Helmke 1987; Karing, Matthäi & Artelt 2011): Bezogen auf Schülermerkmale informiert die *Niveaukomponente* darüber, ob das Lehrendenurteil die klassendurchschnittliche Ausprägungshöhe des interessierenden Merkmals exakt erfasst, überschätzt oder unterschätzt. In der *Streuungskomponente* kommt zum Ausdruck, ob die Lehrperson die Bandbreite bestehender Unterschiede zwischen den Merkmalsträger/innen korrekt einstuft, überzeichnet oder nivelliert. Mithilfe der *Rangordnungskomponente* kann ermittelt werden, wie gut es ihr gelingt, die Merkmalsträger/innen in eine den realen Verhältnissen entsprechende Rangreihe zu bringen, also bspw. treffsicher die besten und schlechtesten Schüler/innen der Klasse zu identifizieren.

Trotz dieses messmethodischen Arsenals lässt sich die individuelle Urteilskraft im Einzelfall nicht absolut bestimmen, weil gemeinhin anerkannte Schwellenwerte fehlen, anhand derer gute von moderaten oder schlechten Diagnoseleistungen trennscharf abgegrenzt werden (z.B. Spinath 2005). Hinlänglich dokumentiert sind dagegen Bedingungsfaktoren *intraindividueller Variationen* (Artelt & Rausch 2014; Südkamp, Kaiser & Möller 2012). So belegen mehrere statusdiagnostische Studien, dass Lehrpersonen die kognitiven und leistungsbezogenen Merkmale ihrer Schüler/innen exakter einschätzen als deren emotional-motivationale Merkmale (z.B. Karing 2009; Urhahne et al. 2010). Überzeugende Hinweise auf eine Fachgebundenheit der Urteilsgüte liefern Lorenz und Artelt (2009), die für wiederholt im Halbjahresabstand erbetene Diagnosen nachweisen können, dass Einschätzungen verschiedener sprachlicher Teilfähigkeiten zu jedem Zeitpunkt untereinander konsistenter ausfallen als Einschätzungen von sprachlichen und mathematischen Teilfähigkeiten. Dass sich die Treffsicherheit von Urteilen auch als Funktion des

verwendeten Akkuratheitsmaßes darstellen lässt, zeigen unter anderem Analysen von Karst (2012) sowie Anders et al. (2010). Sie legen nahe, dass Niveau-, Rang- und Streuungskomponente je unterschiedliche Anforderungen und Fehlerquellen diagnostischer Leistungen implizieren, denen eine Lehrperson nicht gleichermaßen gewachsen sein mag. Darüber hinaus demonstriert Spinath (2005), dass nicht nur verschiedene Akkuratheitsmaße innerhalb eines einzelnen Schülermerkmals, sondern auch gleiche Maße über diverse Merkmale hinweg (namentlich Intelligenz, Fähigkeitsselbstkonzept, Lernmotivation und Leistungsängstlichkeit) mehrheitlich nicht bedeutsam positiv korrelieren.

Insgesamt nähren diese Befunde den Eindruck, dass diagnostische Fähigkeiten *gegenstands- bzw. domänenspezifisch* und *kriteriumsabhängig* ausgeprägt sind und nicht als eindimensionales Konstrukt einer generellen Diagnose(in)kompetenz der betreffenden Lehrperson zu begreifen sind (Schrader 2010). Diese Interpretation legen auch Ergebnisse der SALVE-Studie nahe, in welcher 27 Lehrkräfte lernfortschrittsrelevante Schülermerkmale prozessnah bewerteten, indem sie unmittelbar im Anschluss an eine Mathematikstunde unter anderem beziffern sollten, wie viele Schüler/innen den Stoff der Stunde verstanden hatten und während des Unterrichts aufmerksam und interessiert waren (Hosenfeld, Helmke & Schrader 2002). Übereinstimmungen mit Selbstauskünften, welche die Schüler/innen auf retrospektiven Kurzfragebögen abgaben, fielen dabei recht bescheiden aus. So lag der geschätzte Anteil ‚verständiger' Lernender in rund zwei Dritteln aller abgegebenen Urteile markant unter den Schülerangaben; die Abweichungshöhe betrug im Mittel 16 %, im Extremfall 74 %. Ebenso wurde der Anteil interessierter Lernender mehrheitlich unterschätzt, wenn auch in deutlich geringerem Umfang (durchschnittlich um 6 % und maximal um 48 %).

Ob und wie sehr die Akkuratheit von Urteilen, die eine Lehrperson während des Unterrichts fällt, *situativen Schwankungen* unterliegt, ist bislang nicht untersucht worden. Bei der Suche nach relevanten Grundkategorien situationsbedingter Einflüsse sollten neben *objektivierbaren* Umständen wie etwa beobachtbaren Sozialformen oder Aktivitätsstrukturen auch *subjektive* Erlebensqualitäten oder Zielsetzungen in distinkten, als Situationen definierten, Zeitintervallen berücksichtigt werden (siehe hierzu Beck 1996). Die vorliegende Studie unternimmt erste Schritte auf diesem Weg.

3 Prämissen und Fragestellungen einer explorativen Studie im Berufsschulunterricht

Angesichts des wenig erschlossenen Forschungsfeldes treffen wir zunächst Annahmen über die Schwierigkeiten pädagogischer Prozessdiagnosen und deren mögliche Implikationen für die Urteilsgenauigkeit:

- *Gleichzeitigkeit*: Anders als bei statusdiagnostischen Aufgaben sind Diagnoseleistungen für alle Lernenden und diverse Beurteilungsgegenstände simultan zu erbringen und zu integrieren, was per se hohe Anforderungen an individuelle Informationsverarbeitungskapazitäten stellt (z.B. Krolak-Schwerdt, Böhmer & Gräsel 2009);
- *Spontaneität*: Das laufende Interaktionsgeschehen bietet (methodenabhängig) kaum Rückzugsmöglichkeiten, um in bewussten Reflexionsschleifen abwägende und solide begründbare Urteile zu fällen. Stattdessen müssen unmittelbare Eindrücke des Schülerverhaltens ad hoc in spontane Bewertungen überführt

werden, die ihrerseits flexible didaktische Reaktionen erfordern. Der unausweichliche Handlungsdruck erzwingt somit „hochautomatisierte und schematisierte Zustands-, Veränderungs- und Diskrepanzdiagnosen des Schülerverhaltens, des Unterrichtsverlaufs und der eigenen Handlungseffekte“ (Weinert & Schrader 1986, 11), in denen mentale Skripts und subjektive Erklärungsmodelle unterrichtlicher Ereignisse eine schnelle, aber nicht unbedingt realitätsadäquate Orientierungs- und Handlungssicherheit gewährleisten (Sembill & Seifried 2009; Artelt & Rausch 2014).

- *Beschränkte Beobachtbarkeit lernrelevanter Schülermerkmale bei geringer Validität etlicher Verhaltensindikatoren*: Während kognitive Merkmale wie das aktuelle Verstehen vermutlich relativ sicher aus Fragen und Antworten im Klassengespräch oder der Lösung von Übungsaufgaben erschlossen werden können, existieren für emotional-motivationale Merkmale wie die situative Langeweile nur wenige ökologisch valide Verhaltensindikatoren (Kögler 2015). Damit ist die Urteilsbildung über latente psychische Zustände und deren Veränderungen anhand von oftmals flüchtigen Äußerungen, Aktivitäten und Produkten der Lernenden vor allem im nicht-kognitiven Merkmalsspektrum gefährdet, durch Halo-Effekte und logische Fehlschlüsse verzerrt zu sein (z.B. Helmke 2012; Rausch 2013).
- *Tücken des klasseninternen Referenzrahmens*: Ein bedarfsgerechtes didaktisches Agieren erfordert zumindest eine akkurate Einschätzung mehrheitlich vorhandener Merkmalsausprägungen in der Klassengemeinschaft (z.B. van Ophuysen 2010; Hosenfeld, Helmke & Schrader 2002). Allerdings gilt gerade der klassenbezogene Bewertungsmaßstab als „Achillesferse des Lehrerurteils“ (Lorenz & Artelt 2009, 213). Wo klare kriteriale Bezugsnormen fehlen und Schülervoraussetzungen de facto heterogen ausfallen, mag eine Sortierung der Lernenden entlang der sozialen Vergleichsdimension (Rangkomponente der Diagnosegüte) noch sicher gelingen; Urteile über die absolute Höhe klassendurchschnittlicher Merkmalsausprägungen dürften dagegen besonders fehleranfällig sein und eine Orientierung an dem/der mutmaßlichen ‚Durchschnittsschüler/in’ erheblich erschweren (s. auch Südkamp & Möller 2009).

In einer explorativen Studie haben wir daher Lehrpersonen an beruflichen Schulen mit der Herausforderung klassenbezogener Einschätzungen lernrelevanter Merkmale im laufenden Unterricht konfrontiert und sie gebeten, ihre Urteile mithilfe eines *Continuous-State-Sampling*-Verfahrens bewusst zu fällen. Hiermit verfolgen wir drei Fragestellungen, in denen jeweils die *Niveaukomponente* der Diagnosegüte als Akkuratheitsmaß fungiert.

1. Wie gut können die Lehrpersonen das aktuelle Verstehen sowie die situative Langeweile der Schüler/innen während des Unterrichts einschätzen bzw. in welchem Ausmaß weichen klassenbezogene Diagnosen der Lehrenden von korrespondierenden Selbstauskünften der Lernenden ab?
2. Beeinflussen Facetten des subjektiven Unterrichtserlebens der Lehrpersonen in der jeweiligen Diagnosesituation (z.B. inhaltliche Sicherheit oder Stoffbewältigung) die Akkuratheit ihrer Urteile?
3. Gelingen lehrerseitige Diagnosen des schülerseitigen Verstehens und/oder Langeweileempfindens in bestimmten Aktivitätsstrukturen besser als in anderen? Hängt also die Urteilsgenauigkeit von den in der Diagnosesituation vollzogenen Lehr-Lern-Handlungen ab?

4 Methode

4.1 Stichprobe

Die Stichprobe enstammt einer Studie im Längsschnittdesign, an der zwei Berufsschulklassen teilgenommen haben (vgl. hierzu Golyszny, Kärner & Sembill 2012; Kärner 2015). Bei den 53 Schüler/innen handelt es sich um angehende Industriekaufleute, die über einen Zeitraum von drei Wochen mit je drei zusammenhängenden Stunden pro Woche im Fach ‚Betriebswirtschaftliche Geschäftsprozesse' unterrichtet wurden.

Tab. 1: Stichprobencharakteristik der explorativen Studie an zwei Berufsschulklassen

	Klasse 1	**Klasse 2**
Teilnehmende Schüler/innen	28	25
Ø Alter [Jahre] (M / SD)	20.16 (6.15)	19.21 (3.48)
Männlich	8 (28.6 %)	10 (40 %)
Weiblich	20 (71.4 %)	15 (60 %)

Beide Lehrende – nachfolgend als L1 und L2 bezeichnet – sind männlichen Geschlechts und verfügen über langjährige Berufserfahrung. Dennoch wurden die Analysen zur Diagnosegüte im Unterricht für jede Person separat durchgeführt, um detaillierte Einzelfallbetrachtungen zu ermöglichen.

4.2 Messinstrumente und Operationalisierung

Die prozessbegleitende Untersuchung stützt sich auf zwei Erhebungsmethoden. Mithilfe des *Continuous-State-Samplings* wurden sowohl die Verstehens- und Langeweileangaben der Schüler/innen als auch die korrespondierenden klassenbezogenen Lehrendenurteile erfasst. Dieses Verfahren gewährleistet aufgrund der situationsbezogenen Datenerhebungen eine hohe ökologische Validität, wobei die Abfragen im Unterschied zu klassischen Experience-Sampling-Methoden in festen statt in zufälligen Zeitintervallen erfolgen (vgl. Csíkszentmihályi & Larson 1987; Sembill, Seifried & Dreyer 2008). Als Items dienten in der vorliegenden Studie die Schüleraussagen „*Verstehe, worum es geht*" und „*Mir ist langweilig*" sowie die entsprechenden lehrerseitigen Einschätzungen „*Schüler verstehen, worum es geht*" und „*Schülern ist es langweilig*", deren klassenspezifische Ausprägungen in Abschnitt 5.1 eingehend betrachtet werden. Ebenso wurden auf diesem Wege diverse Erlebensdimensionen der Lehrpersonen im Unterricht erfragt, namentlich die Aspekte „*Fühle mich überfordert*", „*Komme mit dem Stoff durch*", „*Lärmpegel belastet mich*" und „*Bin inhaltlich sicher*". Sämtliche Angaben machten die Probanden auf Mobilen Datenerfassungsgeräten (Palm Tungsten E2®), die hierzu in gleichgetakteten 10-Minuten-Intervallen per Tonsignal aufforderten. Im Verlauf von neun Unterrichtsstunden konnten so 38 Messungen in Klasse 1 und 36 Messungen in Klasse 2 vorgenommen werden. Selbstauskünfte wie auch klassenbezogene Einschätzungen sind stufenlos mit Extremwerten von 0 bis 100 skaliert.

Anhand von *Videoaufzeichnungen* des Unterrichtsgeschehens wurden beobachtbare Aktivitäten im 15-Sekunden-Rhythmus von geschulten Ratern in ein einfaches Klassifikationsschema mit den Kategorien *Schülerarbeitsphasen* und *fragend-entwickelndes Lehrgespräch* eingruppiert. Die Intercoder-Reliabilität wurde für unabhängige Doppelkodierungen über drei Unterrichtsstunden hinweg überprüft und erreicht ein Cohens κ von .73 basierend auf 987 Kodierungen, was gemäß Landis und Koch (1977) als angemessen gilt. Anschließend wurde das Ausmaß der kodierten Aktivitäten in jeder zehnminütigen Unterrichtssequenz, auf die sich die selbst- bzw. klassenbezogenen Auskünfte der Probanden jeweils richteten, ebenfalls minutengenau bestimmt. Um die diagnostische Sensibilität unter verschiedenen unterrichtlichen Bedingungen zu kontrastieren, wurde mit Hilfe prozentualer Anteile an jeder 10-Minuten-Sequenz die jeweils dominante Aktivitätsform ermittelt. Als dominant wurde eine Aktivitätsform eingestuft, sobald sie mit einem Anteil von mindestens 60 % in der betreffenden Sequenz vertreten war. Dementsprechend wurden insgesamt 46 Diagnosesituationen als Lehrgespräch und 23 als Schülerarbeitsphase klassifiziert; 5 Situationen wurden aufgrund uneindeutiger Mischungsverhältnisse von den in Abschnitt 5.3 berichteten Vergleichen ausgeschlossen.

5 Empirische Befunde

5.1 Akkuratheit prozessdiagnostischer Urteile

Die Abbildungen 1 bis 4 stellen den klassenbezogenen Diagnosen der Lehrenden im Verlauf von neun Unterrichtsstunden die klasseninternen Selbsteinschätzungen ihrer Schüler/innen unmittelbar gegenüber. Auffällig ist zunächst, dass das Langeweileempfinden der Lernenden zu jedem Messzeitpunkt (MZP) stark um das jeweilige Klassenmittel streut, welches vor allem in Klasse 2 auch markante Amplituden im Zeitverlauf aufweist (Klasse 1: M = 34.32; SD = 24.06; Klasse 2: M = 37.34; SD = 31.51). Demgegenüber pendelt sich das selbstberichtete Verstehen in beiden Klassen auf einem vergleichsweise homogenen und stabil hohen Durchschnittsniveau ein (Klasse 1: M = 80.54; SD = 20.16; Klasse 2: M = 79.16; SD = 18.73).

Ferner ist ersichtlich, dass beide Lehrenden das mittlere Verstehensniveau ihrer Klasse meist exakter einschätzen als das mittlere Langeweileempfinden, auch wenn ihre Urteile unterschiedliche Fehlertendenzen aufweisen. Während L1 in 34 von 38 MZP das Verstehen um nicht mehr als 16.7 Punkte (d.h. maximal um 19.96 % und durchschnittlich um 6.97 %) verfehlt, zeichnet sich in seinen Diagnosen der Langeweile ein phasenweiser Wechsel von markanten Überschätzungen (um durchschnittlich 27.51 Punkte bei einer Spanne von 1 und 34.8 Punkten) sowie geringen bis moderaten Unterschätzungen (M = -9.3, min. Diff. = -1.31, max. Diff. = -23.64) ab.

L2 überschätzt das mittlere Verstehensniveau der Lernenden mit Ausnahme von drei MZP durchgängig über alle Unterrichtsstunden hinweg (durchschnittlich um 12.82, mindestens um 1.96 und maximal um 26.67 Punkte). Das mittlere Langeweileniveau unterschätzt L2 zunächst monoton während der ersten 21 MZP (M = -22.17, min. Diff. = -7.52, max. Diff. = -37.92), bevor er es während der folgenden 15 Zeitpunkte in rascher, unsystematischer Folge entweder überschätzt, unterschätzt oder aber treffend taxiert.

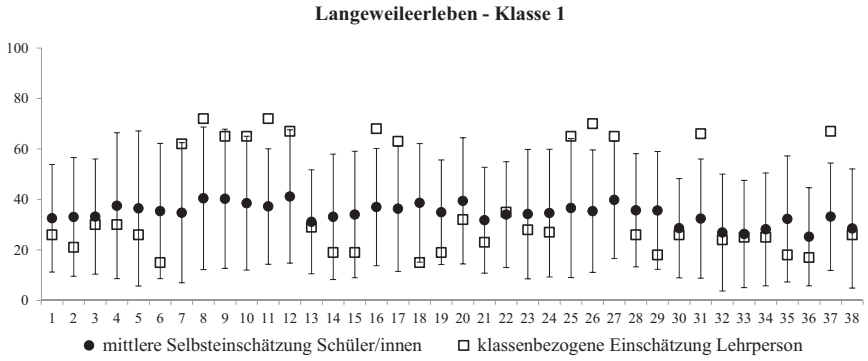

Abb. 1: Akkuratheit prozessdiagnostischer Urteile – Klasse 1 (Langeweile)

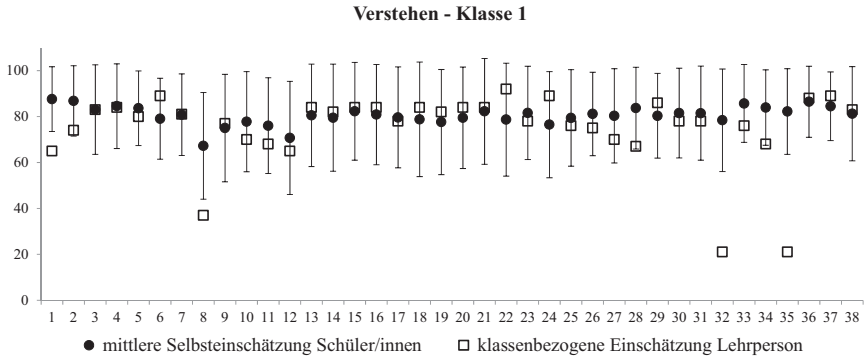

Abb. 2: Akkuratheit prozessdiagnostischer Urteile – Klasse 1 (Verstehen)

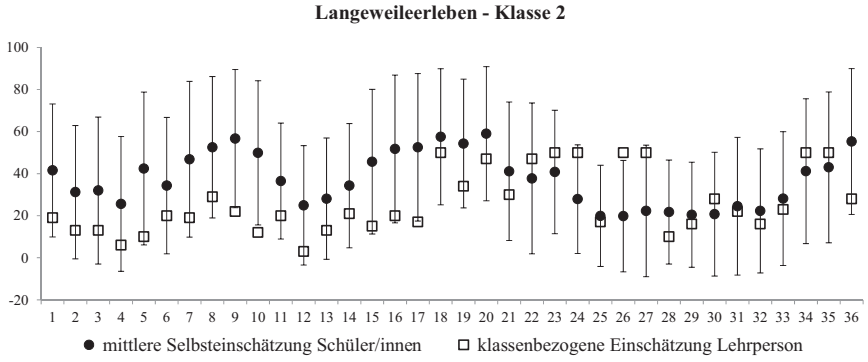

Abb. 3: Akkuratheit prozessdiagnostischer Urteile – Klasse 2 (Langeweile)

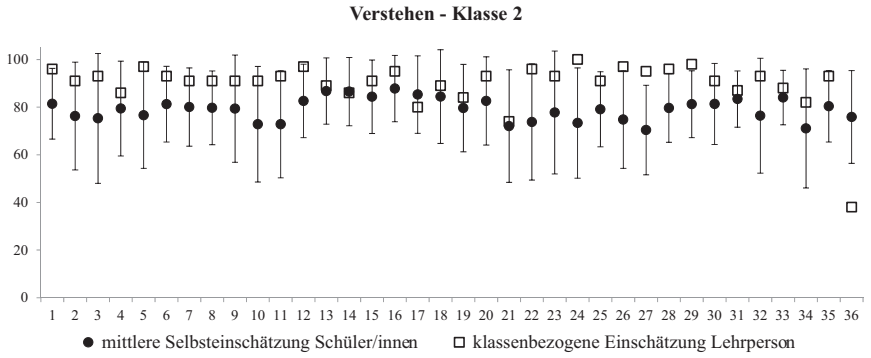

Abb. 4: Akkuratheit prozessdiagnostischer Urteile – Klasse 2 (Verstehen)

5.2 Einflüsse unterrichtsbezogener Erlebensdimensionen der Lehrpersonen auf die Einschätzungsgenauigkeit

Aus Tabelle 2 geht hervor, dass sich das Ausmaß klassenspezifischer Abweichungen zwischen lehrerseitigen Diagnosen und schülerseitigen Angaben zum *aktuellen Verstehen* anhand von unterrichtsbezogenen Erlebensdimensionen der Lehrenden regressionsanalytisch vorhersagen lässt. In Klasse 1 können zwar nur rund 13 %, in Klasse 2 aber knapp ein Viertel der Variabilität in den Differenzwerten durch die in einer Diagnosesituation vorherrschenden Erlebensqualitäten der Lehrperson erklärt werden. Bei beiden Personen stellt der subjektive Eindruck, mit dem Unterrichtsstoff *‚durchzukommen'*, den stärksten Prädiktor dar; bei L2 spielt zudem der wahrgenommene Grad des *momentan herrschenden Lärmpegels* im Klassenraum eine bedeutsame Rolle.

Da die Kriteriumsvariable sowohl positive als auch negative Werte annehmen kann, illustrieren die Abbildungen 5, 6 und 7 die Muster der identifizierten Zusammenhänge. Demnach kann L1 treffsichere Diagnosen (mit Differenzwerten um Null), vor allem dann verbuchen, wenn er das ausgeprägte Empfinden hat, planmäßig im Unterrichtsstoff voranzuschreiten, was wiederum sehr häufig der Fall ist. L2, der generell zur Überschätzung des Schülerverstehens neigt, tut dies umso mehr, je stärker sein Eindruck einer planmäßigen Stoffbewältigung ausgeprägt ist, wobei dieser Eindruck situativ erheblich stärker variiert als bei L1. Interessante Zusatzinformationen liefern daher Zusammenhangsmaße zwischen empfundener Stoffbewältigung und geschätztem Verstehensniveau der Schülern/innen (anstelle der situationsspezifischen Abweichungen von den Lernendenangaben). Sie betragen für L1 $r = .39$ ($p = .016$) sowie für L2 $r = .32$ ($p = .056$) und indizieren in beiden Fällen *positive* Relationen, die bei der Ergebnisinterpretation berücksichtigt werden sollten (siehe hierzu Kapitel 6).

Nur für L2 lässt sich außerdem feststellen, dass seine Diagnosen des aktuellen Verstehens umso mehr mit den Selbstauskünften der Lernenden übereinstimmen, je mehr er sich vom Lärmpegel beeinträchtigt fühlt.

Tab. 2 Einflüsse unterrichtsbezogener Erlebensdimensionen der Lehrpersonen auf die Einschätzungsgenauigkeit

	B	SE (B)	β	p
Konstanter Term				
Lehrperson 1	-24.52	76.26		.750
Lehrperson 2	-3.78	18.51		.840
Fühle mich überfordert				
Lehrperson 1	-.36	.67	-.09	.601
Lehrperson 2	-	-	-	-
Bin inhaltlich sicher				
Lehrperson 1	-1.00	.66	-.24	.136
Lehrperson 2	.11	.21	.08	.603
Komme mit dem Stoff durch				
Lehrperson 1	1.24	.50	.41	.018
Lehrperson 2	.20	.08	.39	.020
Lärmpegel belastet mich				
Lehrperson 1	.13	.47	.04	.781
Lehrperson 2	-.92	.41	-.34	.030

Hinweise: Abhängige Variable: Differenz *Verstehen* (Lehrereinschätzungen - Schüleraussagen), wobei Werte < 0 → Unterschätzung, Werte > 0 → Überschätzung; Der Aspekt *Fühle mich überforder*t wurde bei L2 aufgrund von Multikollinearität ausgeschlossen;
Lehrperson 1: N = 38 MZP. R^2 = .227, korr. R^2 = .134, F(df) = 2.43(4), p = .067
Lehrperson 2: N = 36 MZP. R^2 = .305, korr. R^2 = .240, F(df) = 4.68(3), p = .008

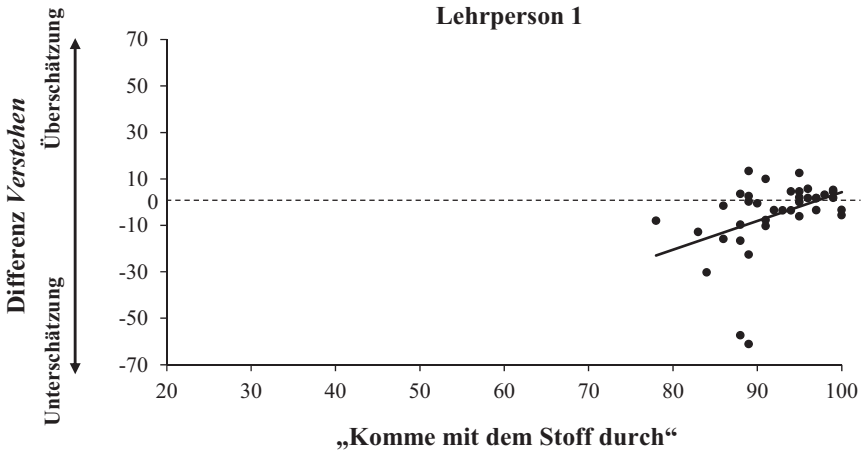

Abb. 5: Differenz „Verstehen“ und „Komme mit dem Stoff durch“ – L1

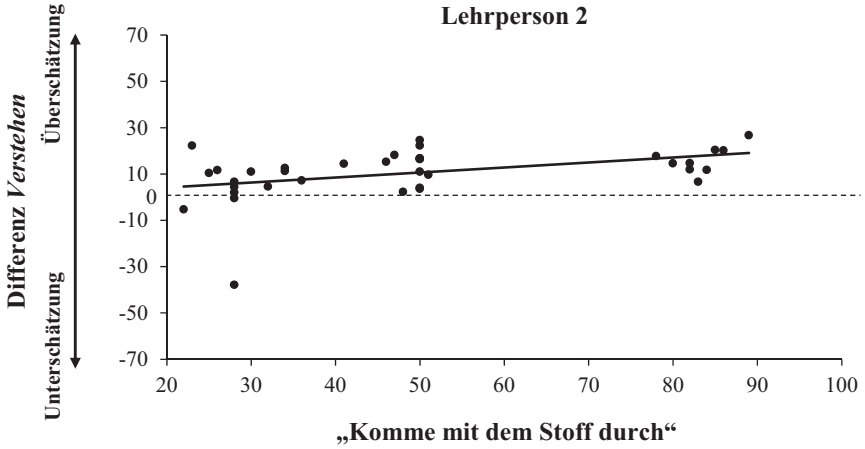

Abb. 6: Differenz „Verstehen“ und „Komme mit dem Stoff durch“ – L2

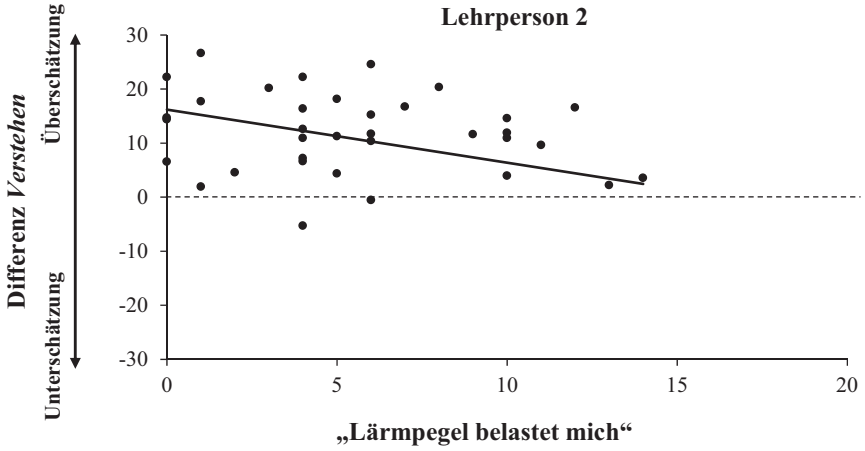

Abb. 7: Differenz „Verstehen“ und „Lärmpegel belastet mich“ – L2

5.3 Abhängigkeit der Einschätzungsgenauigkeit von beobachtbaren Unterrichtsaktivitäten

Abbildung 8 dokumentiert, dass die Einschätzungsgenauigkeit einer jeden Lehrperson auch systematisch davon abhängt, welche unterrichtlichen Aktivitäten schwerpunktmäßig in einer Diagnosesituation durchgeführt werden. L1 erzielt in fragend-entwickelnden Lehrgesprächen (30 Messzeitpunkte) deutlich und kohärent akkuratere Urteile über das aktuelle Verstehensniveau seiner Klasse als in Schülerarbeitsphasen (6 MZP). Für die Diagnosegüte der situativen Langeweile ergeben sich bei L1 im Mittel jedoch keine signifikanten Unterschiede zwischen den

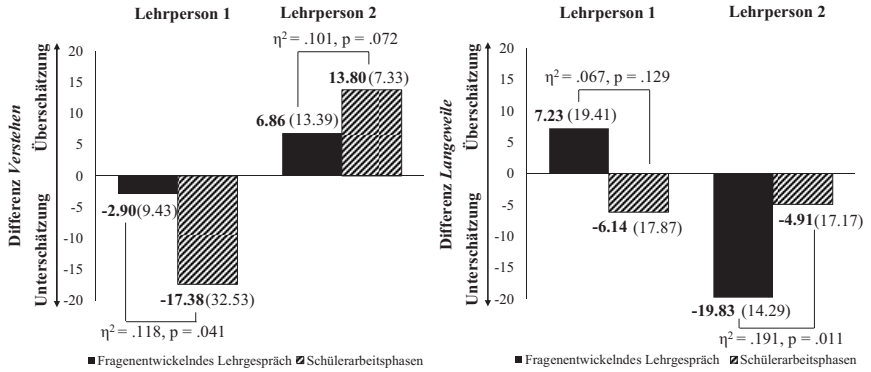

Abb. 8: Abhängigkeit der Einschätzungsgenauigkeit von beobachtbaren Unterrichtsaktivitäten

Aktivitätsformen. Zumindest deutet sich in den durchschnittlichen Abweichungen zwischen Lehrendenurteil und Lernendenangaben an, dass L1 das Langeweileempfinden in Lehrgesprächs-Sequenzen eher überschätzt und in Schülerarbeitsphasen eher unterschätzt.

Auch L2 taxiert das mittlere Verstehen in der Klassengemeinschaft zutreffender, wenn ein Lehrgespräch stattfindet (16 MZP), während die für ihn charakteristische Tendenz zur Überschätzung besonders hoch ausfällt, wenn die Lernenden eigenständig Aufgaben bearbeiten (17 MZP). Das Langeweileniveau der Schülerschaft unterschätzt er hingegen konsequent und teils erheblich in sämtlichen Lehrgesprächs-Sequenzen. In den Schülerarbeitsphasen schwanken seine Urteile zwischen Über- und Unterschätzung, wobei sich in der Durchschnittsbetrachtung wie bei L1 eine leichte Unterschätzung ergibt.

6 Diskussion

Die hier berichtete Untersuchung erfasste die *diagnostische Sensibilität* von Lehrpersonen im laufenden Unterricht mit Hilfe des Continuous-State-Samplings. Den Schwerpunkt der statistischen Auswertungen bildeten die Effekte perzipierter und beobachteter Unterrichtsmerkmale auf die erzielte Urteilsgenauigkeit. Zwar sind die Ergebnisse der durchgeführten Einzelfallanalysen zweier Berufsschullehrkräfte naturgemäß nicht verallgemeinerbar; sie dürfen aber angesichts eines eklatanten Mangels an prozessbegleitenden Erhebungen der Diagnosegüte in natürlichen Settings ‚Pioniercharakter' für sich beanspruchen und können einer hypothesengenerierenden Logik entsprechend durchaus Anregungen für Folgestudien bieten.

6.1 Zur Güte kontinuierlicher Diagnosen lernrelevanter Schülermerkmale im Unterrichtsprozess

Übereinstimmend mit statusdiagnostischen Forschungsbefunden legen unsere prozessdiagnostischen Einzelfallanalysen *gegenstandsabhängig differierende Diagnosefähigkeiten* der Lehrenden, operationalisiert anhand der Niveaukomponente der Diagnosegüte, nahe. Für diese Interpretation spricht, dass klasseninterne Abweichungen zwischen Lehrendenurteil und Lernendenangaben zum *aktuellen Verstehen* in 68 % aller erfassten Diagnosesituationen geringer ausfallen als für die *situative Langeweile*.

Ob es sich hierbei um eine ‚gute' oder ‚schlechte' Diagnoseleistung handelt, ist aufgrund fehlender Richt- und Schwellenwerte in der wissenschaftlichen Literatur nicht pauschalisierend zu beantworten. Immerhin lässt sich anhand der Gegenüberstellung gegenstandsspezifischer Einschätzungsgenauigkeiten vermuten, dass Lehrpersonen ein laufendes Monitoring des klassendurchschnittlichen Verstehensniveaus relativ leicht gelingt, eine kontinuierliche Beurteilung des mittleren Langeweilelevels hingegen eher schwerfällt. Die faktisch hohe Variabilität des schülerindividuellen Langeweileempfindens sowohl innerhalb einer Klassengemeinschaft als auch über den Verlauf der betrachteten neun Unterrichtsstunden hinweg mag zur vergleichsweise geringen Treffsicherheit in diesem Gegenstandsbereich ebenso beitragen wie das weitgehende Fehlen eindeutig dechiffrierbarer Verhaltensindikatoren für emotional-motivationale Zustände der Lernenden.

Des Weiteren ist davon auszugehen, dass sich die Lehrpersonen unserer Studie aufgrund ihrer langjährigen Berufserfahrung der möglichen Streubreiten schülerseitiger Selbstzuschreibungen von aktuellem Verstehen und situativer Langeweile durchaus bewusst sind. Daher stellt sich die Anschlussfrage, ob und welche Ankerpunkte prozessbegleitende Diagnosen in der alltäglichen Unterrichtspraxis *außerhalb oder anstelle* des geschätzten Klassenmittels finden, um pädagogische Handlungen zu regulieren. Forschungsarbeiten zur Lehrer-Schüler-Interaktion legen nahe, dass vorrangig eine Orientierung an besonders leistungsstarken Schüler/innen erfolgt (z.B. Lipowsky, Rakoczy, Pauli, Reusser & Klieme 2007).

Darüber hinaus ließe sich grundsätzlich fragen, in welcher Weise prozessdiagnostische Leistungen von Lehrkräften im Unterrichtsalltag gewinnbringend unterstützt werden könnten, um die individuelle Treffsicherheit zu erhöhen und adäquate didaktische Justierungen vornehmen zu können (s. auch Sloane 2012). Der Einsatz von *Classroom Response Systemen* mag hierfür Potenziale bieten, die es jedoch noch empirisch auszuloten gilt (Howe & Knutzen 2012).

6.2 Zum Einfluss perzipierter und beobachteter Unterrichtsmerkmale auf die Akkuratheit der Lehrendenurteile

Signifikante Effekte *unterrichtsbezogener Erlebensdimensionen* der Lehrpersonen auf die Diagnosegüte lassen sich regressionsanalytisch nur für den Gegenstandsbereich des aktuellen Verstehens nachweisen. In konsistenter Weise scheinen beide Lehrpersonen dazu zu neigen, *subjektive Eindrücke eigener Stoffbewältigung* in einer Diagnosesituation auf das aktuelle Verstehen ihrer Schüler/innen zu *projizieren*: Je optimistischer sie selber sind, gut mit dem Unterrichtsstoff ‚durchzukommen', desto

eher unterstellen sie auch den Lernenden, dem aktuellen Unterrichtsgeschehen gut folgen zu können. Während allerdings L2 mit *wachsender* Zuversicht, planmäßig voranzuschreiten, das mittlere Verstehensniveau in seiner Klasse *noch stärker überschätzt* als er dies ohnehin stets tut, *büßt* L1 seine zumeist hohe Treffsicherheit mit *sinkender Zuversicht* ein Stück weit *ein*. Dieser Befund lässt es sinnvoll erscheinen, in Untersuchungen zur situativen Variabilität der Einschätzungsgenauigkeit grundlegende individuelle Fehlertendenzen bei der Urteilsbildung als Kontrollvariablen zu berücksichtigen. Zudem sind sowohl aus einer statistischen als auch aus einer konzeptionellen Blickrichtung wahrscheinliche Rückkoppelungseffekte zwischen den hier erfassten Variablen zu beachten. So dürften bspw. diagnostizierte Verstehens*defizite* bei den Schüler/innen ihrerseits nachfolgende Urteile zur Stoffbewältigung während einer Unterrichtsstunde negativ beeinflussen, weil sie Stützmaßnahmen wie etwa wiederholende Erläuterungen erforderlich machen oder zumindest unerwartet lange Überlegungs-, Antwort- und Bearbeitungszeiten in laufenden Gesprächs- oder Arbeitsphasen bedingen.

Ein weiterer auffälliger Befund besteht darin, dass sich die Diagnosegüte bei dem generell zur Überschätzung tendierenden Probanden mit zunehmendem Grad eines als störend empfundenen Lärmpegels in der Diagnosesituation bedeutsam *verbessert*. Hieraus ließe sich die These gewinnen, dass Lehrpersonen eine steigende Unruhe in der Klasse als greifbaren Indikator sinkenden Verstehens – und infolgedessen abschweifender Nebengespräche oder irritierter Nachfragen bei dem/der Banknachbar/in – werten, welcher zur Korrektur unrealistisch hoher Niveauschätzungen genutzt wird. In der Unterrichtspraxis dürften sich im Gefolge solcher zunächst noch diffusen Wahrnehmungen von zunehmender Unruhe im Klassenzimmer gerichtete Aufmerksamkeits- und Suchstrategien anschließen, um Störquellen genauer zu lokalisieren und gegenzusteuern. Zur detaillierteren Analyse dieser kritischen Phasen kann es sich in weiteren triangulativen Auswertungsschritten lohnen, einzelne Sequenzen des videografierten Unterrichtsgeschehens zu inspizieren, die unmittelbar nach deutlichen Ausschlägen in den kontinuierlich berichteten Erlebensqualitäten der Lehrpersonen folgen.

Dass auch *objektiv variierende Klassenaktivitäten* mit der Exaktheit prozessdiagnostischer Urteile in Beziehung stehen, geht aus der Kontrastierung von Schülerarbeitsphasen und fragend-entwickelnden Lehrgesprächen hervor. Im Einklang mit unseren theoretischen Vorüberlegungen gelingt es beiden Lehrpersonen im *direkten dialogischen Austausch* mit den Lernenden besser, deren *aktuelles Verstehen* einzuschätzen, als wenn Lerninhalte eigenständig oder in Kleingruppen erarbeitet werden (siehe zu den Potenzialen qualitativ hochwertiger Lehrgespräche auch Bauer-Klebl 2010). Dennoch lässt ein feinkörnigeres Analyseraster in Folgestudien auch differentielle Ergebnisse innerhalb des Bereichs der Schülerarbeitsphasen erwarten. Ob sich eine Lehrperson in solchen Phasen maximal zurücknimmt, ob sie umhergeht, um unauffällig zu beobachten und auf Anfrage wohldosierte Anregungen zu geben, oder ob sie gar mit gezielten Fragen und konkreten Anweisungen in die Aufgabenbearbeitung interveniert (vertiefend Seifried & Klüber 2006; Link 2011), könnte unmittelbare Auswirkungen auf ihre Einschätzungsgenauigkeit des aktuellen Verstehensniveaus haben.

Überzufällige aktivitätsbedingte Unterschiede in der Treffsicherheit von *Langeweile-Einschätzungen* zeigen sich wiederum nur für eine der beiden Lehrpersonen, fallen bei ihr aber markant im Sinne einer durchgängigen Unterschätzung des Lan-

geweileempfindens der Schüler/innen im fragend-entwickelnden Lehrgespräch aus. Hieraus ergibt sich die weiterführende Frage, inwieweit individuelle Sichtweisen auf verschiedene Unterrichtsmethoden die Urteilsfindung über das schülerseitige Erleben grundlegend prägen.

Literatur

Anders, Y., Kunter, M., Brunner, M., Krauss, S. & Baumert, J. (2010). Diagnostische Fähigkeiten von Mathematiklehrkräften und ihre Auswirkungen auf die Leistungen ihrer Schülerinnen und Schüler. *Psychologie in Erziehung und Unterricht*, 57(3), 175–193.

Artelt, C. & Rausch, T. (2014). Accuracy of Teacher Judgments. When and for What Reasons? In S. Krolak-Schwerdt, S. Glock & M. Böhmer (Eds.*), Teachers' Professional Development: Assessment, Training, and Learning*, (pp. 27–43). Rotterdam, Boston & Taipei: Sense Publishers.

Bauer-Klebl, A. (2010). Interaktionsprozesse im Lehrgespräch – Lernchance oder Zeitverschwendung? In R. Nickolaus, G. Pätzold, H. Reinisch & T. Tramm (Hrsg.), *Handbuch Berufs- und Wirtschaftspädagogik*, (S.122–126). Bad Heilbrunn: Julius Klinkhardt.

Beck, K. (1996). Die „Situation" als Bezugspunkt didaktischer Argumentationen – Ein Beitrag zur Begriffspräzisierung. In W. Seyd & R. Witt (Hrsg.), *Situation, Handlung, Persönlichkeit. Kategorien wirtschaftspädagogischen Denkens*, (S. 87–98). Hamburg: Feldhaus.

Beck, E., Baer, M., Guldimann, T., Bischoff, S., Brühwiler, C., Müller, P., Niedermann, R., Rogalla, M. & Vogt, F. (2008). Adaptive Lehrkompetenz. Analyse von Struktur, Veränderbarkeit und Wirkung handlungssteuernden Lehrerwissens. Münster, New York, München & Berlin: Waxmann.

Behrmann, L. & Souvignier, E. (2013). The Relation Between Teachers' Diagnostic Sensitivity, their Instructional Activities, and their Students' Achievement Gains in Reading. *Zeitschrift für Pädagogische Psychologie*, 27(4), 283–293.

Breuer, K. (2010). Leistungsbewertung und Unterrichtsevaluation. In R. Nickolaus, G. Pätzold, H. Reinisch & T. Tramm (Hrsg.), *Handbuch Berufs- und Wirtschaftspädagogik*, (S. 195–201). Bad Heilbrunn: Julius Klinkhardt.

Brühwiler, C. (2014). Adaptive Lehrkompetenz und schulisches Lernen: Effekte handlungssteuernder Kognitionen von Lehrpersonen auf Unterrichtsprozesse und Lernergebnisse der Schülerinnen und Schüler. Münster: Waxmann.

Csíkszentmihályi, M. & Larson, R. (1987). Validity and Reliability of the Experience-Sampling Method. *The Journal of Nervous and Mental Disease*, 175(9), 526–536.

Dünnebier, K., Gräsel, C. & Krolak-Schwerdt, S. (2009). Urteilsverzerrungen in der schulischen Leistungsbeurteilung. *Zeitschrift für Pädagogische Psychologie*, 23(3–4), 187–195.

Golyszny, K., Kärner, T. & Sembill, D. (2012). Unischulprojekt „Belastung und Stress am Arbeitsplatz Schule, insbesondere in Lehr-Lern-Kontexten" – Relevanz der Thematik und erste Ergebnisse. *Wirtschaft und Erziehung*, 64(7), 221–224.

Helmke, A. (2012). Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts. Seelze-Velber: Klett-Kallmeyer.

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*, (S. 65–82). Weinheim & Basel: Beltz.

Howe, F. & Knutzen, S. (2012). Digitale Medien in der gewerblich-technischen Berufsausbildung. Eine Expertise im Auftrag des Bundesinstituts für Berufsbildung am Beispiel der Einsatzmöglichkeiten digitaler Medien in Lern-und Arbeitsaufgaben. Bremen, Hamburg.

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

Karing, C. (2009). Diagnostische Kompetenz von Grundschul- und Gymnasiallehrkräften im Leistungsbereich und im Bereich Interessen. *Zeitschrift für Pädagogische Psychologie*, 23(3–4), 197–209.

Karing, C., Matthäi, J. & Artelt, C. (2011). Genauigkeit von Lehrerurteilen über die Lesekompetenz ihrer Schülerinnen und Schüler in der Sekundarstufe I – Eine Frage der Spezifität? *Zeitschrift für Pädagogische Psychologie*, 25(3), 159–172.

Kärner, T. (2015). Erwartungswidrige Minderleistung und Belastung im kaufmännischen Unterricht. Analyse pädagogischer, psychologischer und physiologischer Aspekte. Frankfurt a. M.: Lang.

Karst, K. (2012). Kompetenzmodellierung des diagnostischen Urteils von Grundschullehrern. Münster, New York, München & Berlin: Waxmann.

Klauer, K. J. (2014). Formative Leistungsdiagnostik: Historischer Hintergrund und Weiterentwicklung zur Lernverlaufsdiagnostik. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (S. 1–18). Göttingen: Hogrefe.

Klieme, E. & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik*, 54(2), 222–237.

Kögler, K. (2015). Langeweile in kaufmännischen Unterrichtsprozessen – Entstehung und Wirkung emotionalen Erlebens ungenutzter Zeitpotentiale. Frankfurt a.M., Bern, Bruxelles, New York, Oxford, Warszawa & Wien: Lang.

Kögler, K., Bauer, C. & Sembill, D. (2011). Auf dem Weg zur Selbstorganisation – Wochenplanarbeit in Unterrichtsprozessen der Wirtschaftsschule. In K. Wilbers (Hrsg.), *Die Wirtschaftsschule – Verdienste und Entwicklungsperspektiven einer bayerischen Schulart*, (S. 329–346). Aachen: Shaker.

Krolak-Schwerdt, S. Böhmer, M. & Gräsel, C. (2009). Verarbeitung von schülerbezogenen Informationen als zielgeleiteter Prozess. Der Lehrer als „flexibler Denker". *Zeitschrift für Pädagogische Psychologie*, 23(3–4), 175–186.

Landis, J. R. & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159–174.

Link, F. (2011). Problemlöseprozesse selbstständigkeitsorientiert begleiten. Wiesbaden: Vieweg+Teubner Research.

Lipowsky, F., Rakoczy, K., Pauli, C., Reusser, K. & Klieme, E. (2007). Gleicher Unterricht – gleiche Chancen für alle? Die Verteilung von Schülerbeiträgen im Klassenunterricht. *Unterrichtswissenschaft*, 35(2), 125–147.

Lorenz, C. & Artelt, C. (2009). Fachspezifität und Stabilität diagnostischer Kompetenz von Grundschullehrkräften in den Fächern Deutsch und Mathematik. *Zeitschrift für Pädagogische Psychologie*, 23(3–4), 211–222.

McElvany, N., Schroeder, S., Hachfeld, A., Baumert, J., Richter, T., Schnotz, W., Horz, H. & Ullrich, M. (2009). Diagnostische Fähigkeiten von Lehrkräften bei der Einschätzung von Schülerleistungen und Aufgabenschwierigkeiten bei Lernmedien mit instruktionalen Bildern. *Zeitschrift für Pädagogische Psychologie*, 23(3–4), 223–235.

Oser, F., Heinzer, S. & Salzmann, P. (2010). Die Messung der Qualität von professionellen Kompetenzprofilen von Lehrpersonen mit Hilfe der Einschätzung von Filmvignetten: Chancen und Grenzen des advokatorischen Ansatzes. *Unterrichtswissenschaft*, 38(1), 5–28.

Rausch, T. (2013). Wie Sympathie und Ähnlichkeit Leistungsbeurteilungen beeinflussen kann. *Erziehung & Unterricht*, 163(9–10), 937–944.

Ricken, G. (2006). Lernprozessdiagnostik. In K.-H. Arnold, U. Sandfuchs & J. Wiechmann (Hrsg.), *Handbuch Unterricht*, (S. 639–642). Bad Heilbrunn: Julius Klinkhardt.

Schrader, F.-W. (2010). Diagnostische Kompetenz von Eltern und Lehrern. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie*, (S. 102–108). Weinheim & Basel: Beltz.

Schrader, F.-W. (2013). Diagnostische Kompetenz von Lehrpersonen. *Beiträge zur Lehrerbildung*, 31(2), 154–165.

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

Seidel, T. & Prenzel, M. (2007). Wie Lehrpersonen Unterricht wahrnehmen und einschätzen – Erfassung pädagogisch-psychologischer Kompetenzen mit Videosequenzen. *Zeitschrift für Erziehungswissenschaft*, 10 (Sonderheft 8), 201–216.

Seifried, J. (2004). Fachdidaktische Variationen in einer selbstorganisationsoffenen Lernumgebung. Eine empirische Untersuchung im Rechnungswesenunterricht. Wiesbaden: Deutscher Universitäts-Verlag.

Seifried, J. & Klüber, C. (2006). Lehrerinterventionen beim selbstorganisierten Lernen. In P. Gonon, F. Klauser & R. Nickolaus (Hrsg.), *Bedingungen beruflicher Moralentwicklung und beruflichen Lernens*, (S. 153–164). Wiesbaden: VS Verlag für Sozialwissenschaften.

Sembill, D. & Seifried, J. (2009). Konzeptionen, Funktionen und intentionale Veränderungen von Sichtweisen. In O. Zlatkin-Troitschanskaia, K. Beck, D. Sembill, R. Nickolaus & R. Mulder (Hrsg.), *Lehrprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung*, (S. 345–354). Weinheim & Basel: Beltz.

Sembill, D., Seifried, J. & Dreyer, K. (2008). PDAs als Erhebungsinstrument in der beruflichen Lehr-Lern-Forschung – Ein neues Wundermittel oder bewährter Standard? *Empirische Pädagogik*, 22(1), 64–77.

Sloane, P. F. E. (2012): Dr. House als Chefdidaktiker – Diagnose und Unterricht: Welche Diagnostik benötigen Lehrende? *Zeitschrift für Berufs- und Wirtschaftspädagogik*, 108(2), 161–168.

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

Südkamp, A. & Möller, J. (2009). Referenzgruppeneffekte im Simulierten Klassenraum: direkte und indirekte Einschätzungen von Schülerleistungen. *Zeitschrift für Pädagogische Psychologie*, 23(3–4), 161–174.

Südkamp, A., Kaiser, J. & Möller, J. (2012). Accuracy of Teachers' Judgments of Students' Academic Achievement: A Meta-Analysis. *Journal of Educational Psychology*, 104(3), 743–762.

Tenberg, R. (2012). Lerndiagnostik im kompetenzorientierten Unterricht. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, 108(4), 481–490.

Urhahne, D., Zhou, J., Stobbe, M., Chao, S.-H., Zhu, M. & Shi, J. (2010). Motivationale und affektive Merkmale unterschätzter Schüler. Ein Beitrag zur diagnostischen Kompetenz von Lehrkräften. *Zeitschrift für Pädagogische Psychologie*, 24(3–4), 275–288.

van Buer, J. & Zlaktin-Troitschanskaia, O. (2009). Diagnostische Lehrerexpertise und adaptive Steuerung unterrichtlicher Entwicklungsangebote. In J. van Buer & C. Wagner (Hrsg.), *Qualität von Schule – Ein kritisches Handbuch*, (S. 381–400). Frankfurt a. M., Berlin, Bern, Bruxelles, New York, Oxford & Wien: Lang.

van Ophuysen, S. (2010). Professionelle pädagogisch-diagnostische Kompetenz – eine theoretische und empirische Annäherung. In N. Berkemeyer, W. Bos, H. G. Holtappels, N. McElvany & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung. Daten, Beispiele und Perspektiven*, (S. 203–234). Weinheim & München: Juventa.

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

Wuttke, E. & Seifried, J. (2013). Diagnostic competence of (prospective) teachers in vocational education. An Analysis of Error Identification in Accounting Lessons. In K. Beck & O. Zlatkin-Troitschanskaia (Eds.), *From Diagnostics to Learning Success. Proceedings in Vocational Education and Training*, (pp. 225–240). Rotterdam, Boston & Taipei: Sense Publishers.

Anschrift der Autoren/innen: Dr. Julia Warwas, Otto-Friedrich-Universität Bamberg, Lehrstuhl für Wirtschaftspädagogik, Kärntenstr. 7, 96052 Bamberg. Mail: julia.warwas@uni-bamberg.de
Dr. Tobias Kärner, Otto-Friedrich-Universität Bamberg, Lehrstuhl für Wirtschaftspädagogik, Kärntenstr. 7, 96052 Bamberg. Mail: tobias.kaerner@uni-bamberg.de
M.Sc. Klaudia Golyszny, Otto-Friedrich-Universität Bamberg, Lehrstuhl für Wirtschaftspädagogik, Kärntenstr. 7, 96052 Bamberg. Mail: klaudia.golyszny@uni-bamberg.de