VERBESSERTE URTEILSGENAUIGKEIT UND ADÄQUATERE SCHÜLERFÖRDERUNG DURCH DEN PSYCHOMETRISCHEN HUNTER-SCHMIDT-ANSATZ

Esther Kaufmann*, Werner W. Wittmann**

**Pädagogische Hochschule Zentralschweiz (Zug)*
Institut für Bildungsmanagement und Bildungsökonomie
Zugerbergstrasse 3
CH-6300 Zug
esther.kaufmann@phz.ch

***Otto-Selz-Institut für Angewandte Psychologie*
Universität Mannheim
L13, 15, Raum 125
D-68131 Mannheim
wittmann@tnt.psychologie.uni-mannheim.de

Schlagworte: *Lehrerurteile, Urteilsgenauigkeit, Psychometrische Meta-Analyse*

Zusammenfassung. *Die Urteilsgenauigkeit bei Aufgaben zur Beurteilung von Leistungs- oder Lerninteresse von Schülern im Rahmen der Sozialen Urteilstheorie wurde bisher nur mit „Bare bones"-Meta-Analysen evaluiert, obwohl mit psychometrischen Meta-Analysen zusätzliche Artefakte korrigiert werden können. Unsere psychometrische Meta-Analyse zeigt deutlich, dass die Urteilsgenauigkeit im Bildungskontext mit psychometrischen Korrekturen verbessert werden kann. Zur Verbesserung der Urteilsgenauigkeit im Bildungskontext sollten aber auch psychometrisch korrigierte Expertenmodelle zur Unterstützung bei der Urteilsbildung herangezogen werden. Durch die verbesserte Urteilsgenauigkeit im Bildungskontext mittels des psychometrischen Hunter-Schmidt-Ansatzes profitieren auch die Schüler, die eine adäquatere Förderung erhalten, wenn sie genauer beurteilt werden.*

---

1. Einleitung

Eine wichtige Tätigkeit des Lehreralltages ist die Beurteilung und Förderung von Schülern (siehe Artelt & Gräsel, 2009; Helmke, Hosenfeld, & Schrader, 2004). Lehrpersonen beurteilen, ob Schüler Texte verstehen oder ob Schüler tatsächlich motiviert sind, zu lernen. Die Beurteilung von Schülern geht mit der geeigneten Förderung einher. Ungenaue Urteile von Lehrpersonen können zu einer Unter- oder Überforderung von Schülern führen. Eine Überforderung kann zur Überbeanspruchung, zu langfristigen Belastungen und zu Stress führen. Unterforderte Schüler langweilen sich im Unterricht und erbringen keine optimalen Leistungen.

Zusammengefasst können ungenaue Urteile von Lehrpersonen für die Schulausbildung sowie für die nachfolgende Berufswahl der Schüler negative Folgen haben. Es ist daher sehr wichtig, dass Urteile von Lehrpersonen genau sind. Dadurch stellt sich die Frage, wie genau diese Urteile sind. Kann die Urteilsgenauigkeit von Lehrpersonen verbessert werden?

Im Rahmen der Sozialen Urteilstheorie (Hammond, Stewart, Brehmer, & Steinmann, 1975), welche auf Brunswik's Forschung (1952) basiert, kann die Urteilsgenauigkeit evaluiert und die Gründe für ihre Ungenauigkeit können aufgedeckt werden. Dazu werden mehrere Urteile von mindestens einer Person anhand eines Kriteriums verglichen. Zum Beispiel werden in der Studie

von Cooksey, Freebody und Davidson (1986) 18 Lehrpersonen hinsichtlich des Textverstehens von 118 Schülern befragt. Diese Urteile wurden anhand eines Kriteriums, des tatsächlichen Textverstehens der Schüler, durch einen Textverständnis-Test ermittelt und verglichen. Somit kann mittels des Zusammenhangs (Korrelation) zwischen den Lehrerurteilen zum Textverstehen der Schüler und dem tatsächlichen Textverstehen der Schüler die Genauigkeit der Lehrerurteile bestimmt werden. Im Rahmen der Sozialen Urteilstheorie können die Gründe der Urteilsgenauigkeit oder -ungenauigkeit auch durch die Tucker'sche Linsengleichung (1964) aufgedeckt werden. Die Tucker'sche Linsengleichung wurde 1964 von Hammond (Hammond, Hursch, & Todd, 1964) sowie Hursch (Hursch, Hammond, & Hursch, 1964) initiiert, aber von Tucker (1964) entscheidend modifiziert. Seit dieser Modifikation ist die ursprüngliche Linsengleichung unter dem Namen Tucker'sche Linsengleichung bekannt (siehe Gleichung 1):

$$r_a = GR_sR_e + C\sqrt{1-R_e^2}\sqrt{1-R_s^2} \qquad (1)$$

Mit der Tucker'schen Linsengleichung kann die Urteilsgenauigkeit von einzelnen Personen oder Gruppen (r_a) durch den Umweltfaktor (R_e), den Personenfaktor (R_s) oder durch den Zusammenhang zwischen dem Umwelt- und dem Personenfaktor bestimmt werden (G, C). In Bezug auf unser Beispiel aus der Studie von Cooksey et al. (1986) kann nun folgendes bezüglich der Urteilsgenauigkeit der Lehrperson ermittelt werden:

1. Inwieweit die Urteilsgenauigkeit der Lehrperson durch den Umweltfaktor der Aufgabe (R_e) bestimmt wurde. Ist zum Beispiel die Beurteilung des Textverstehens von Schülern überhaupt genau beurteilbar? Sind die verwendeten Urteilsinformationen, wie z.B. Geschlecht des Schülers und Lesegeschwindigkeit, brauchbar, um das Textverstehen des Schülers zu beurteilen? Inwieweit kann aufgrund der verwendeten Informationen das Textverstehen von Schülern beurteilt werden?
2. Der zweite Faktor, die Personenkomponente, zeigt auf, inwieweit die Urteilsgenauigkeit auf die Lehrperson zurückzuführen ist. Beurteilt die Lehrperson das Textverstehen von unterschiedlichen Schülern konsistent, d.h., benutzt sie immer die gleichen Informationen zur Urteilsbildung, oder bevorzugt die Lehrperson Schüler einer bestimmten Nation, eines bestimmten Geschlechts? Inwieweit sind daher die Urteile einer Lehrperson konsistent?
3. Der letzte Faktor weist auf den Zusammenhang zwischen den bereits erwähnten Faktoren, Person (R_s) und Aufgabe (R_e), hin. Inwieweit hängen diese Faktoren linear (G) oder nicht-linear (C) zusammen?

Seit der Entwicklung der Tucker'schen Linsengleichung wurde diese in zahlreichen Studien (siehe Karelaia & Hogarth, 2008, für einen allgemeinen Überblick) und in unterschiedlichen Kontexten, wie bei medizinischen und wirtschaftlichen Urteilen, angewendet. Auch im Bildungskontext wurde die Tucker'sche Linsengleichung verwendet, um die Güte der Urteile zu evaluieren (siehe Kaufmann & Athanasou, 2009; Kaufmann, Sjödahl, & Mutz, 2007).

Gleichzeitig wurde die Tucker'sche Linsengleichung auch weiterentwickelt (siehe Castellan, 1972; Stewart, 1976) und benutzt, um Expertenmodelle (Camerer, 1981; Karelaia & Hogarth, 2008) zu bilden. Expertenmodelle können zur Unterstützung bei Urteilen herangezogen werden und somit auch die Urteilsgenauigkeit verbessern. Im Rahmen der Sozialen Urteilstheorie wird der Erfolg von Expertenmodellen wie folgt berechnet (siehe Gleichung 2, Camerer, 1981, S. 413):

$$\Delta = GR_e - r_a \qquad (2)$$

Der Erfolg von Expertenmodellen wird durch die Subtraktion des Expertenmodells von der Urteilsgenauigkeit der urteilenden Person (r_a, siehe oben) berechnet. Das Expertenmodell wird aus den zwei Komponenten lineare Wissenskomponente (G, siehe oben) und der Umweltkomponente (R_e, siehe oben) der Tucker'schen Linsengleichung gebildet.

Da die Tucker'sche Linsengleichung auch in Studien im Bildungskontext angewendet wurde, stellen sich folgende Fragen:

1. Wie genau sind Urteile im Bildungskontext (siehe Tucker'sche Linsengleichung: r_a)?
2. Was sind die zugrunde liegenden Faktoren für ungenaue Urteile (siehe Komponenten der Tucker'schen Linsengleichung: R_e, R_s, G, C)?
3. Können Expertenmodelle die Urteilsgenauigkeit im Bildungskontext tatsächlich verbessern (siehe Gleichung 2)?

Da viele Studien die Tucker'sche Linsengleichung verwendeten um die Urteilsgenauigkeit zu bestimmen, sind bereits Meta-Analysen publiziert worden (siehe Karelaia & Hogarth, 2008; Kaufmann & Athanasou, 2009). Diese bisherigen Publikationen bestimmten im Rahmen der Sozialen Urteilstheorie die Urteilsgenauigkeit über Personen hinweg. Diese Meta-Analysen benutzten auch den sogenannten „Bare bones"-Ansatz, obwohl eine psychometrische Meta-Analyse genauer ist, da bei diesen noch zusätzliche Artefakte korrigiert werden (siehe Hunter & Schmidt, 2004; Schmidt, 2010; Wittmann, 1985, 2009). Unser Beitrag evaluiert die Urteilsgenauigkeit im Rahmen der Sozialen Urteilstheorie mittels einer psychometrischen Meta-Analyse nach Hunter und Schmidt (2004) und vergleicht diese anschliessend mit den Ergebnissen einer herkömmlichen „Bare bones"-Meta-Analyse. Die folgende Gleichung (3) gibt einen Überblick über die unterschiedlichen Artefakt-Korrekturen, welche mit dem Hunter-Schmidt-Ansatz korrigiert werden können. Zu erwähnen ist, dass es sich bei dieser Gleichung (3) um eine Erweiterung (Wittmann, 1985, 2009) des Hunter-Schmidt-Ansatzes handelt, da zusätzlich auch die Symmetrie (d.h. Generalitätsniveau der beiden Variablen Urteil und Kriterium, welche sich am Besten entsprechen sollten) berücksichtigt wird.

Aus der Schätzung der Urteilsgenauigkeit mit dem Hunter-Schmidt-Ansatz nach Wittmann (1985, 2009) wird deutlich, dass bei einer Nichtberücksichtigung der einzelnen Artefakte sechs Gefahren der Unter- sowie zwei Gefahren der Überschätzung der eigentlichen Urteilsgenauigkeit bestehen. Im Folgenden wird die Korrektur der Artefakte berücksichtigt, um Unter- und Überschätzungen der Urteilsgenauigkeit im Bildungskontext zu vermeiden (siehe auch Kaufmann, 2010, S. 16ff.).

$$r_{a,\ \textit{wahre Urteils-genauigkeit}} = S\sqrt{r_{tt}^{Rs}\, r_{tt}^{Re}}\; G\, R_s\, R_e + e \qquad (3)$$

- S: Selektionseffekte — 1 Gefahr der Überschätzung, 1 Gefahr der Unterschätzung aufgrund von Streuungsreduktion oder -erweiterung
- $r_{tt}^{Rs}\, r_{tt}^{Re}$: Psychometrische Reliabilität der Urteile und des Evaluationskriteriums — 2 Gefahren der Unterschätzung
- $G\, R_s\, R_e$: Personen sowie Umweltfaktor (Konstruktreliabilität) — 2 Gefahren der Unterschätzung (Mangel der Symmetrie)
- e: Stichprobenfehler — 1 Gefahr der Überschätzung (positiver Fehler); 1 Gefahr der Unterschätzung (negativer Fehler)

Es gibt 6 Gefahren der Unterschätzung im Vergleich zu 2 Gefahren der Überschätzung der wahren Urteilsgenauigkeit

2. Methode

2.1 Literatursuche und Codierung der Studien

Die ausführliche Literatursuche beschränkt sich auf die Artikel von 1964 bis August 2008. Dabei wurden fünf unterschiedliche Strategien verwendet:

1. Die Mailing-List der Brunswik Society wurde über das Projekt informiert.
2. Einschlägige Studien wurden im Brunswik Society Newsletter (1991-2007) gesucht.
3. In Datenbanken (z.B. ERIC, PSYNDEX) wurde mit Schlagwörtern (z.B. Urteil, Soziale Urteilstheorie) nach weiterer Literatur gesucht.
4. In Online-Datenbanken (z.B. Google, Google Scholar) wurde mit denselben Schlagwörtern und einer anschliessenden Referenzsuche gesucht.
5. In einschlägigen Büchern (z.B. Hammond & Stewart, 2001) und Artikeln (z.B. Tucker, 1964) wurde nach Referenzen gesucht.

Für weitere Informationen über die Literatursuche sowie nachfolgende Kontrollstrategien, Ausschlusskriterien (d.h. Urteilsgenauigkeit vs. Lern- und Feedbackstudien) und die Codierung der einzelnen Studien verweisen wir auf Kaufmann (2010, S. 38ff.). Unsere ausführliche Literatursuche ergab 31 Studien, wovon drei Studien dem Bildungskontext zuzuordnen sind (Athanasou & Cooksey, 2001; Cooksey et al., 1986; Wiggins & Kohen, 1971).

2.2 Studien

Die drei Studien im Bildungskontext enthalten vier Urteilsaufgaben, welche im Folgenden entsprechend ihrem Publikationsjahr beschrieben werden.

In der ersten Publikation (Wiggins & Kohen, 1971) beurteilten 98 Psychologiestudenten 110 Profile. Diese Profile enthielten zehn Informationen von angehenden Studenten. Die Urteilsaufgabe bestand darin, die *Leistungen* der angehenden Studenten anhand des Notendurchschnittes im ersten Universitätsjahr zu beurteilen. Die Urteilsgenauigkeit wurde durch den Zusammenhang zwischen den tatsächlichen Noten der Studenten nach einem Jahr und den Urteilen bestimmt.

In der nachfolgenden Studie von Cooksey et al. (1986) beurteilten jeweils 20 Lehrer 118 Profile von Kindergarten-Kindern. Die Lehrer beurteilten aufgrund von jeweils fünf Informationen der Kinder erstens das *Textverstehen* und zweitens den Umfang des *Wortschatzes*. Die Urteilsgenauigkeit wurde durch den Zusammenhang der Lehrerurteile mit den Werten in einem Textverständnis-Test sowie einem Wortschatz-Test bestimmt.

In der letzten Studie von Athanasou und Cooksey (2001) beurteilten 18 angehende Erziehungswissenschaftler die Profile von 120 Studenten hinsichtlich deren *Lerninteresses*. Die Profile der Studenten bestanden aus 20 Informationen. Die Urteilsgenauigkeit wurde durch den Zusammenhang zwischen dem tatsächlichen Interesse der Studenten anhand eines Tests und den Urteilen der angehenden Erziehungswissenschaftlern ermittelt.

Alle drei Studien enthielten die Angaben der einzelnen Komponenten der Tucker'schen Linsengleichung, obwohl diese nicht auf der individuellen Ebene (d.h. idiographischer Ansatz) in der Studie von Wiggins und Kohen (1971) berichtet wurden.

2.3 Psychometrischer Hunter-und-Schmidt-Ansatz

Um einen Aggregations-Fehler (Robinson, 1950) zu vermeiden, wird zuerst der idiographische Ansatz, die Urteilsgenauigkeit von einzelnen Personen, dargestellt.

Danach werden die einzelnen Komponenten der Tucker'schen Linsengleichung mittels einer „Bare bones"-Meta-Analyse im Vergleich zu einer psychometrischen Meta-Analyse dargestellt, um deren Verbesserungspotenzial aufzuzeigen (siehe Hunter & Schmidt, 2004; Schmidt, 2010; Wittmann, 1985, 2009, siehe auch Gleichung 3). Zusätzlich werden die Ergebnisse mittels Robustness-Analysen (z.B. Publikationsfehler, unterschiedliche Gewichtungsstrategien sowie unterschiedliche Korrekturstrategien) überprüft.

Anschliessend werden die korrigierten Komponenten der Tucker'schen Linsengleichung verwendet, um ein Expertenmodell zu bilden (siehe Gleichung 3). Für ausführliche Informationen verweisen wir auf Kaufmann (2010, S. 83 ff.).

3. Resultate

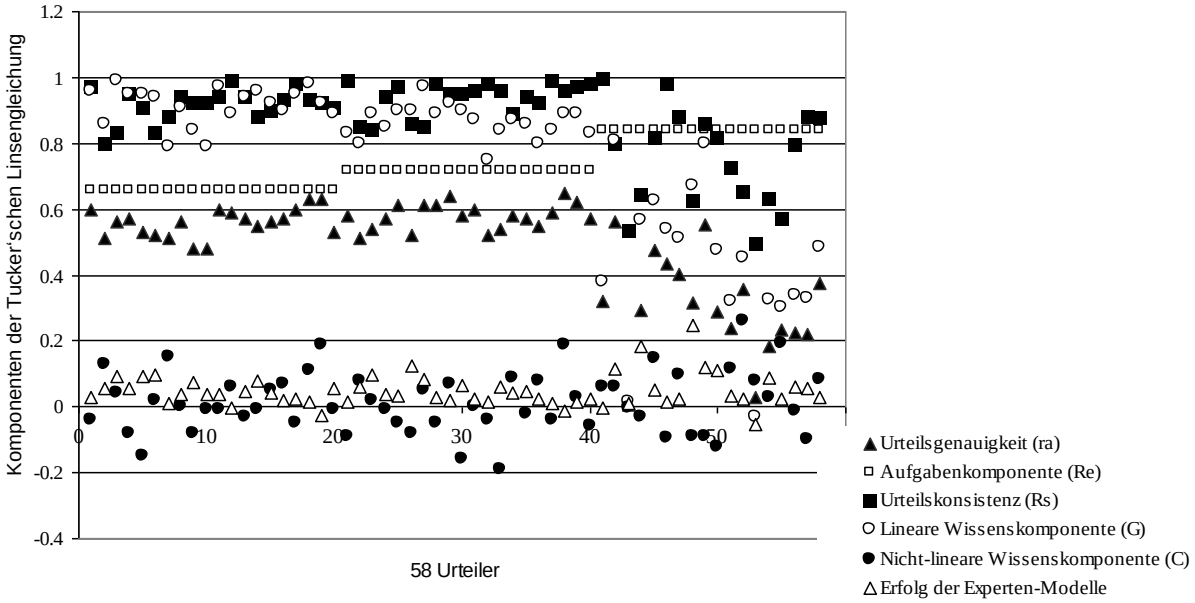
3.1 Idiographische Analyse

In Grafik 1 sind die Komponenten der Tucker'schen Linsengleichung von 58 einzelnen Urteilenden dargestellt, welche in zwei Studien vorhanden waren (Athanasou & Cooksey, 2001; Cooksey et al., 1986). Die einzelnen Komponenten der Tucker'schen Linsengleichung umfassen drei Aufgaben. Alle drei Aufgabenkomponenten – Textverstehen (.66), Wortschatz (.72) sowie Lerninteresse (.84) – sind hoch und können daher auch von den Lehrpersonen genau beurteilt werden.

Die Urteilsgenauigkeit in den drei Aufgaben variiert bei den einzelnen Urteilenden von minimal .05 in der Aufgabe zum Lerninteresse zu maximal .65 bei einem Urteilenden, der den Wortschatz von Schülern beurteilte. Diese grosse Heterogenität der Urteilsgenauigkeit ist auch in der Grafik 1 ersichtlich und widerspiegelt, dass die gleichen Aufgaben sehr unterschiedlich beurteilt werden.

Die Urteilskonsistenz ist hoch über alle Aufgaben hinweg. Die niedrigste Urteilskonsistenz (.49) liegt bei einem Urteilenden, der das Lerninteresse einschätzte, und die höchste (.99) bei einem Urteilenden, der das Textverstehen beurteilte. Die lineare Wissenskomponente (G_{MAX} = .99) ist vergleichbar mit der Urteilskonsistenz sehr hoch, obwohl es einzelne Ausreisser gibt, wie die lineare Wissenskomponente (-.03) eines Urteilenden, der das Lerninteresse von Schülern beurteilt hatte. Im Vergleich zur linearen Wissenskomponente beinhaltet die nicht-lineare Wissenskomponente in diesen drei Aufgaben nur niedrige Werte (C_{MAX} = .19, C_{MIN} = .26).

Obwohl die Urteilsgenauigkeit sehr heterogen ist, zeigt sich deutlich, dass die Expertenmodelle eine Verbesserung der Urteilsgenauigkeit herbeiführen können, da der Erfolg von nur fünf Expertenmodellen gleich oder unter Null ist, was bedeutet, dass in diesen Fällen die Modelle die Urteilsgenauigkeit nicht verbessern können. Es ist dabei zu berücksichtigen, dass die Erfolgseinschätzung der Expertenmodelle konservativ und daher zu Ungunsten der Modelle ausfällt, da bisher noch keine Artefakte korrigiert wurden (siehe Gleichung 3). Bei diesen fünf Urteilenden, bei welchen das Expertenmodell nicht erfolgreich war, beurteilte einer die Wortschatzaufgabe (Δ = .00) und jeweils zwei beurteilten das Textverstehen (Δ = -.03, -.01) oder das Lerninteresse (Δ = .00, -.05). Werden die einzelnen Komponenten der Tucker'schen Linsengleichung bei diesen fünf Urteilenden verglichen, fällt auf, dass die zwei Urteilenden, welche das Lerninteresse einschätzten, eine niedrige Urteilsgenauigkeit (r_a = .32, .03) aufwiesen, die auf einer niedrigen linearen Wissenskomponente (G = .38, -.03) basiert.



Grafik 1: Die Komponenten der Tucker'schen Linsengleichung bei 58 einzelnen Urteilenden

Für einen Vergleich der Komponenten der Tucker'schen Linsengleichung von einzelnen Urteilenden im Bildungskontext mit einzelnen Urteilenden in anderen Kontexten (z.B. Medizin, Wirtschaft) verweisen wir auf Kaufmann et al. (2007) oder Kaufmann (2010).

Nachdem im Folgenden die Komponenten der Tucker'schen Linsengleichung für jede Aufgabe einzeln beschrieben werden, werden aufgrund der Heterogenität der Komponenten der Tucker'schen Linsengleichung die nachfolgenden Meta-Analysen zeigen, ob diese Heterogenität auf Artefakten basiert.

3.2 Psychometrische Hunter-Schmidt-Analyse

In der Tabelle 1 sind die einzelnen Komponenten der Tucker'schen Linsengleichung über die einzelnen Aufgaben nach Genauigkeit der Urteile geordnet. Obwohl sich die Urteilsgenauigkeit in den einzelnen Aufgaben unterscheidet, ist in jeder dieser vier Aufgaben das Expertenmodell erfolgreich und führt somit zu einer Verbesserung der Urteilsgenauigkeit.

	Komponenten der Tucker'schen Linsengleichung					
Aufgaben	r_a	R_e	R_s	G	C	Δ
Wortschatz	.58	.72	.94	.87	.03	.05
Textverstehen	.56	.66	.91	.92	.02	.05
Leistung	.33	.69	.65	.72	.01	.17
Lerninteresse	.31	.84	.76	.44	.10	.06

Tabelle 1: Die Komponenten der Tucker'schen Linsengleichung in den einzelnen Bildungskontext-Aufgaben

In der Tabelle 1 sind die Komponenten der Tucker'schen Linsengleichung für jede der vier Urteilsaufgaben gemittelt. Es ist ersichtlich, dass die nicht-lineare *C*-Komponente nur einen kleinen Anteil ausmacht, d.h., die lineare Komponente widerspiegelt die Urteilsbildung besser als nicht-lineare Komponente. Dies weist auf die Bildung von linearen anstelle von nicht-linearen Expertenmodellen hin (CR_e). Die Urteilsgenauigkeit ist in zwei Aufgaben nur mässig. Die mässige

Urteilsgenauigkeit bei der Beurteilung von Lerninteresse basiert auf einer nur mässigen linearen Komponente. Daher könnte bei dieser Aufgabe die Urteilsgenauigkeit durch die verbesserte lineare Integration der Information gesteigert werden.

Eine nachfolgende „Bare bones"-Meta-Analyse korrigiert die über die vier Aufgaben gemittelten Komponenten der Tucker'schen Linsengleichung hinsichtlich der Anzahl der Urteilenden (Stichprobenfehler, siehe Gleichung 3). Die „Bare bones"-Meta-Analyse ist in der Tabelle 2 jeweils der ersten Zeile bei den einzelnen Komponenten der Tucker'schen Linsengleichung zu entnehmen.

Komponente		*r*	var_{korr}	80% Cl		75%
r_a		.39	.00	.39	.39	178
	Psychometrisch	.51	.00	.51	.51	355
R_e		.70	.00	.70	.70	257
	Psychometrisch	.74	.00	.74	.74	257
R_s		.73	.01	.69	.88	44*
	Psychometrisch	.93	.00	.93	.93	554
G		.73	.01	.60	.86	36*
	Psychometrisch	.97	.00	.97	.97	453
C		.02	.00	.02	.02	3000
	Psychometrisch	.03	.00	.03	.03	3348

Anmerkungen: r = Komponenten der Tucker'schen Linsengleichung. var_{korr} = Varianz des wahren Wertes. *80% Cl* = Credibility Interval. *75%* = Prozentanteil der beobachteten Varianz bezüglich aller Artefakte; ist diese unter 75%, dann sind Moderator-Variablen angezeigt. *Varianz kann mit den bisherigen Artefaktkorrekturen nicht erklärt werden.

Tabelle 2: Vergleich der „Bare bones"-Meta-Analyse mit der psychometrischen Meta-Analyse anhand der vier Aufgaben im Bildungskontext

Die „Bare bones"-Meta-Analyse zeigt eine moderate Urteilsgenauigkeit über die vier Aufgaben hinweg auf, obwohl die vier Urteilsaufgaben gut beurteilbar wären. Diese moderate Urteilsgenauigkeit ergibt sich hauptsächlich wegen der hohen Urteilskonsistenz und der linearen Wissenskomponente. Lassen sich diese ersten Ergebnisse anhand einer psychometrischen Meta-Analyse bestätigen? Eine nachfolgende psychometrische Analyse korrigiert weitere Artefakte wie Messfehler. Die Ergebnisse sind in der Tabelle 2 jeweils unter den Ergebnissen der „Bare bones"-Meta-Analyse in der zweiten Zeile der Komponenten der Tucker'schen Linsengleichung aufgelistet.

Die Urteile über unterschiedliche Aufgaben im Bildungskontext sind nun alle hoch (r_a = .51, var_{corr} = .00, N = 156, k = 4), wie auch der Personen- (R_s = .99) und Aufgabenfaktor (R_e = .73). Unsere Evaluation zeigt, dass Bildungsaufgaben gut beurteilt werden können und dass Lehrpersonen bei diesen Aufgaben eine konsistente Urteilsstrategie anwenden. Zusätzlich ist der Zusammenhang zwischen Aufgaben- und Lehrpersonen linear (G = .97) und nur zu einem kleinen Teil nicht-linear (C = .03).

Für weitere Robustness-Analysen verweisen wir auf Kaufmann (2010).

Zusammengefasst zeigt unsere psychometrische Analyse, dass alle Komponenten der Tucker'schen Linsengleichung ohne Schätzung durch eine psychometrische Meta-Analyse

eindeutig unterschätzt werden. Ebenso wird durch weitere Artefaktkorrekturen die Varianz der einzelnen Komponenten der Tucker'schen Linsengleichung reduziert.

Aufgrund der erhaltenen artefakt-korrigierten Komponenten der Tucker'schen Linsengleichung lassen sich auch die Expertenmodelle korrigiert berechnen – GR_e – .72 anstatt ohne Korrektur .51, daher steigert sich auch der Erfolg von Expertenmodellen von .12 auf .33. Dadurch schlussfolgern wir, a) dass der Erfolg von Expertenmodellen bisher unterschätzt wurde, da die Modelle keine Artefaktkorrektur beinhalteten, und b) dass artefakt-korrigierte Expertenmodelle im Bildungskontext erfolgreich eingesetzt werden können.

4. Diskussion und Schlussfolgerungen

Unsere psychometrische Meta-Analyse nach Hunter-Schmidt (2004) verbesserte durch die Artefaktkorrektur die Urteilsgenauigkeit (r_a = .51) im Bildungskontext. Zusätzlich zeigen die psychometrisch korrigierten Faktoren der Tucker'schen Linsengleichung und deren Verwendung in Expertenmodellen eindeutig, dass noch bessere, genauere Expertenmodelle – auch bereits mit Daten von nur einer Lehrperson – modelliert werden können. Psychometrisch korrigierte Expertenmodelle sollten daher auch im Bildungskontext vermehrt angewendet werden, da sie die Urteilsgenauigkeit deutlich verbessern (Δ = .33). Dabei ist zu beachten, dass die hohen Werte der linearen im Vergleich zu nicht-linearen Komponenten auf die Verwendung von linearen Modellen (Dawes, 1971) hinweisen. Die verbesserte Urteilsgenauigkeit durch psychometrische Expertenmodelle verringert auch bei Selektionen, wie Übertrittsprüfungen, die falsche Positiv- (β-Fehler) und die negative Positiv-Rate (α-Fehler). Durch diese verbesserte Urteils- und Selektionsgenauigkeit werden Schülerinnen und Schüler adäquater gefördert und somit werden auch mögliche Folgeeffekte einer falschen Selektion, wie die Unter- oder Überforderung von Schülerinnen und Schülern, reduziert.

Zusätzlich bestätigt unser Beitrag, dass die Urteilsgenauigkeit ohne Korrekturen im Bildungskontext sehr heterogen ist (siehe Hoge & Coladarci, 1989), obwohl die vier beschriebenen Aufgaben genau beurteilt werden könnten (R_e = .70). Unsere Studie ist auch im Einklang mit bisherigen Studien, welche genauere Urteile bei der Beurteilung von Schülerleistungen als bei emotional-motivationalen Aufgaben zeigten (siehe Karing, 2009, S. 199).

Mit der herkömmlich verwendeten „Bare bones"-Meta-Analyse würde man nur eine mäßige, moderate Urteilsgenauigkeit (r_a = .39) im Bildungskontext annehmen. Daher wird bisher die Urteilsgenauigkeit im Bildungskontext, welche im Rahmen der Sozialen Urteilstheorie mittels einer „Bare bones"-Meta-Analyse evaluiert wurde, deutlich unterschätzt. Diese moderate Urteilsgenauigkeit ist aber vergleichbar mit der aggregierten Urteilsgenauigkeit in anderen Kontexten (z.B. Medizin, Wirtschaft), welche auch moderat ist (siehe Karelaia & Hogarth, 2008; Kaufmann & Athanasou, 2009).

Bei unserer Meta-Analyse ist zu kritisieren, dass nur eine kleine Anzahl von Studien – und diese nur im Rahmen der Sozialen Urteilstheorie – berücksichtigt wurde. Zukünftige psychometrische Meta-Analysen sollten daher auch weitere Studien (siehe Hoge & Coladarci, 1989) zur Urteilsgenauigkeit beinhalten. Zusätzlich ist u.a. zu beachten, dass nur Testwerte als Evaluationskriterium in den Studien verwendet wurden. In künftigen Analysen sollten auch andere Kriterien, wie z.B. eine zweite Expertenmeinung (siehe Kaufmann & Wittmann, 2009) sowie das bisher vernachlässigte Konzept der Symmetrie (siehe Wittmann, 1985, 2009, siehe auch spezifische vs. global Konzept, Helmke et al., 2004), berücksichtigen werden, um die genaue Urteilsfähigkeit von Lehrpersonen und den Erfolg von Expertenmodellen im Bildungskontext nicht nur im Rahmen der Sozialen Urteilstheorie zu bestätigen.

5. Bibliografie

Quellenangaben, welche mit einem Stern (*) markiert sind, wurden bei den vorgestellten Meta-Analysen berücksichtigt.

Artelt, C. & Gräsel, C. (2009). Diagnostische Kompetenz von Lehrkräften. *Zeitschrift für Pädagogische Psychologie, 23*(3-4), 157-160.

*Athanasou, J. A. & Cooksey, R. W. (2001). Judgment of factors influencing interest: An Australian study. *Journal of Vocational Education Research, 26*(1), 77-96.

Brunswik, E. (1952). The conceptual framework of psychology. *International Encyclopedia of Unified Science.* Chicago: University of Chicago Press.

Camerer, C. (1981). General conditions for the success of bootstrapping models. *Organizational Behavior and Human Performance, 27*(3), 411-422.

Castellan, N. J. (1972). The analysis of multiple criteria in multiple-cue judgment tasks. *Organizational Behavior and Human Performance, 8*(2), 242-261.

*Cooksey, R. W., Freebody, P. & Davidson, G. R. (1986). Teachers' predictions of children's early reading achievement: An application of social judgment theory. *American Educational Research Journal, 23*(1), 41-64.

Dawes, R. M. (1974). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571-582.

Hammond, K. R., Hursch, C. J. & Todd, F. J. (1964). Analyzing the components of clinical inference. *Psychological Review, 71*(6), 438-456.

Hammond, K. R. & Stewart, T. R. (Eds.). (2001). *The essential Brunswik: Beginnings, explications, applications*. Oxford, UK: University Press.

Hammond, K. R., Stewart, T. R., Brehmer, B. & Steinmann, D. O. (1975). Social judgment theory. In M. F. Kaplan & S. Schwartz (Eds.), *Human judgment and decision processes* (pp. 271-317). New York: Academic Press, Inc.

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

Hoge, D. H. & Coladarci, T. (1989). Judgments of academic achievement: A review of literature. *Review of Educational Research, 59*(3), 297-313.

Hunter, J. E. & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Hursch, C. J., Hammond, K. R. & Hursch, J. L. (1964). Some methodological considerations in multiple-cue probability learning studies. *Psychological Review, 71*(1), 42-60.

Karelaia, N. & Hogarth, R. (2008). Determinants of linear judgment: A meta-analysis of lens studies. *Psychological Bulletin, 134*(3), 404-426.

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

Kaufmann, E. (2010). *Flesh on the bones: A critical meta-analytical perspective of achievement lens studies* (Doctoral dissertation: MADOC: University of Mannheim). Retrieved from: http://madoc.bib.uni-mannheim.de/madoc/volltexte/2010/2892

Kaufmann, E., & Athanasou, J. A. (2009). A meta-analysis of achievement as defined by the lens model equation. *Swiss Journal of Psychology, 68,* 1-14. doi: 10.1024/1421-0185.68.2.99

Kaufmann, E., Sjödahl, L. & Mutz, R. (2007). The idiographic approach in social judgment theory: A review of components of the lens model equation components. *International Journal of Idiographic Science, 2*.

Kaufmann, E. & Wittmann, W. W. (2009). *Do we underestimate the validity of linear expert models?* Poster presented to the Society for Judgment and Decision Making (SJDM), Boston (MA), November, 22, 2009.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*(2), 351-357.

Schmidt, F. (2010). Detecting and correcting the lies that data tell. *Perspectives on Psychological Science, 5*(3), 233-242.

Stewart, T. R. (1976). Components of correlations and extensions of the lens model equation. *Psychometrika, 41*(1), 101-120.

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond and Hursch and by Hammond, Hursch and Todd. *Psychological Review, 71*(6), 528-530.

*Wiggins, N. & Kohen, E. S. (1971). Man versus model of man revisited: The forecasting of graduate school success. *Journal of Personality and Social Psychology, 19*(1), 100-106.

Wittmann, W. W. (1985). *Evaluationsforschung: Aufgaben, Probleme und Anwendungen.* Berlin: Springer-Verlag.

Wittmann, W. W. (2009). Evaluationsmodelle. In H. Holling (Ed.). *Enzyklopädie der Psychologie. Themenbereich B Methodologie und Methoden. Serie IV Evaluation – Band 1. Grundlagen und statistische Methoden der Evaluationsforschung* (pp. 59-98). Göttingen: Hogrefe.