

## **Instruktionssensitivität als Aspekt gültiger Testwertinterpretation in der Erfassung der Wirksamkeit von Unterricht**

Die empirische Erfassung der Wirksamkeit von Schule und Unterricht mit dem Ziel, (a) die Grundlagen des Lehrens und Lernens zu verstehen, (b) das Bildungssystem zu beschreiben und weiterzuentwickeln, sowie (c) Handlungsempfehlungen für die pädagogische Praxis abzuleiten, die das Lernen von Schüler:innen verbessern, ist ein Kernanliegen empirischer Bildungsforschung (z.B. Fend, 1991; Köller, 2011). Insbesondere Ergebnisse aus Leistungsmessungen bei Schülerinnen und Schülern dienen dabei regelmäßig als ein zentrales Kriterium für die Beurteilung der Wirksamkeit von Unterricht (Klieme, 2018). Gültige Rückschlüsse über die Wirksamkeit des Unterrichts auf Basis der Testwerte von Schülerinnen und Schülern setzen instruktionssensitive Tests voraus; also Tests, die potentielle Effekte des Unterrichts erfassen können (z. B. Popham, 2007; Polikoff, 2010).

Zwar wird Instruktionssensitivität in der Regel angenommen, jedoch selten empirisch überprüft. Dadurch bleibt im Falle des Ausbleibens erwarteter Unterrichtseffekte unklar, ob (a) der eingesetzte Test nicht instruktionssensitiv oder (b) der untersuchte Unterricht nicht effektiv war. Zudem können verschiedene Tests zur Messung des gleichen Konstrukts sehr unterschiedlich instruktionssensitiv sein und entsprechend zu teils widersprüchlichen Rückschlüssen über die Wirksamkeit desselben Unterrichts führen (Grossman, Cohen, Ronfeldt, & Brown, 2014; Naumann, Rieser, Musow, Hochweber, & Hartig, 2019).

Meine Arbeit zielt darauf ab, in solchen Situationen passende empirische Evidenz für die valide Testwertinterpretation bereitzustellen, wenn Rückschlüsse über die Wirksamkeit von Unterricht anhand schulischer Leistungsmessungen gezogen werden sollen. Im Vortrag stelle ich zunächst einen psychometrischen Rahmen für die Messung der Instruktionssensitivität von Items vor, um danach einen Einblick in zwei laufende Forschungsarbeiten zu zwei Fragen zu geben: (a) wie wirkt sich die Selektion von Items anhand ihrer Instruktionssensitivität auf die Sensitivität des resultierenden Tests aus und (b) wie lassen sich zielgerichtet instruktionssensitive Leistungstests konstruieren?