Institut zur Qualitätsentwicklung
im Bildungswesen

Prof. Dr. Olaf Köller, Humboldt-Universität zu Berlin,
Institut zur Qualitätsentwicklung im Bildungswesen

Berlin, 14.11.2006

Stellungnahme zum Text von Joachim Wuttke: "Fehler, Verzerrungen, Unsicherheiten in der PISA-Auswertung"

Wuttke stellt in seinem Aufsatz die Gültigkeit der PISA-Ergebnisse in vielfältiger Weise in Frage. Ich möchte nur einige der Argumente kommentieren.

1.) Von eher geringer Bedeutung sind seine Vorwürfe gegen die **Verzerrungen der verschiedenen Länderstichproben**. Er argumentiert, dass die Stichproben einzelner Länder auf Grund geringeren Beteiligungsquoten und fehlender Repräsentativität in Hinblick auf Geschlecht und Alter verzerrt seien und das Ranking der Länder ungenau und auch ungerecht sei. Hier skizziert Wuttke ein Problem, das in internationalen Studien nicht erst seit PISA existiert. Als Konsequenz werden Länder, welche die Zielvorgaben nicht erreichen, aus dem internationalen Vergleich ausgeschlossen oder markiert (mit Fußnote). Mängel in den länderspezifischen Stichproben im nationalen Vergleich haben in PISA 2000 zum Ausschluss von Hamburg und Berlin geführt. Insofern muss man den PISA-Autoren zugute halten, dass sie sehr sensibel mit dem Problem umgehen.

2.) Abhängigkeit von Beteiligungsquoten. Wuttke argumentiert, dass **Beteiligungsquoten und Leistungen korrelieren**: Je höher die Beteiligungsquoten, desto geringer die Leistungen im Mittel. Wie schon zu Punkt 1 ist zu sagen, dass man in PISA dieser Problematik Rechnung trägt, dass Länder mit zu geringen Beteiligungsquoten vom Vergleich ausgeschlossen werden.

3.) Wuttke kritisiert **differentielle Beteiligungsquoten in den Schülerfragebögen** und stellt fest, dass die Testleistungen der Schüler, die keinen Fragebogen ausfüllen, unter denen liegen, die einen Bogen bearbeitet haben. Auch dies Problem ist bekannt, die dadurch entstehenden Verzerrungen sind sehr gering, dies stellt auch Wuttke fest, das Deutsche PISA-Konsortium schätzt in der Regel die fehlenden Werte, wodurch Verzerrungen drastisch reduziert werden.

4.) Die Kernkritik richtet sich allerdings gegen die **Skalierung der Leistungsdaten**. Wuttke kritisiert, dass die von ACER verwendete Software vermutlich nicht einmal den nationalen Projektpartnern vorgelegt wird und er auf Grund der Dokumentationen die Skalierung nicht nachvollziehen könne. Hier sei ACER kläglich gescheitert. Mit eigenen Programmen könne er aber die Ergebnisse nachrechnen und komme zu anderen Befunden. Hier entlarvt sich Wuttke als Laie, der weder in der Lage ist, gescheit zu recherchieren, noch die Fachdiskussion zu suchen und zu lesen. Das von ACER in PISA verwendete Softwarepaket ist Conquest, eine Skalierungssoftware, die jeder Mensch auf der Welt, auch Herr Wuttke, käuflich erwerben kann. Was Conquest genau leistet, wie man zu optimalen Schätzungen der Item- und Personenparameter kommt, ist im Conquest-Handbuch sehr gut dokumentiert. Conquest gilt aktuell als eine äußerst leistungsstarke Software, in der nur Schätzverfahren verwendet werden, die State of the Art sind. Kaufen und lesen statt zu spekulieren hätte hier Wuttke

weitergeholfen. Die Validität der Schätzungen in Conquest kann man leicht feststellen, wenn man sie mit Schätzungen aus anderen kommerziell erhältlichen Programmen vergleicht. Hier zeigt sich, dass die Schätzungen identisch sind, es also keinen Grund gibt anzunehmen, dass Conquest in PISA falsche Lösungen generiert hat. Eigene Skalierungen des Deutschen PISA-Konsortiums in 2000 haben immer die Ergebnisse von ACER replizieren können. Hätte Herr Wuttke sich die Mühe gemacht, sich mit Conquest zu beschäftigen, hätte er einfach herausgefunden, wie das genaue Prozedere in PISA war und wie einfach man auf die Metrik 500/100 gekommen ist.

5.) Wuttke argumentiert, dass das verwendete 1-Parameter-Raschmodell ungeeignet ist, um Item- und Personenparameter zu schätzen. Hier hätten Mehr-Parameter-Modelle verwendet werden müssen. Auch hier erweist sich Wuttke als Laie. Hätte er sich mit der großen Literaturmenge zu IRT-Modellen auseinandergesetzt, wäre er zu anderen Schlüssen gekommen. Bos hat die IGLU-2001-Daten, die mit dem 3-Parameter-Modell skaliert worden sind, noch einmal mit Conquest (1-Parameter-Modell) skaliert und dabei festgestellt, dass sich die Item- und Personenparameter quasi nicht unterschieden, die unterschiedlichen Modelle hatten keinen differenziellen Effekt auf das Kompetenzmodell. In der IRT-Literatur ist man sich einig, dass die verschiedenen Modelle „unter dem Strich“ zu weitgehend identischen Schätzungen der Personenparameter führen und es eher eine Frage der Weltanschauung ist, welches man verwendet (Europa und Australien eher das 1-Parameter-Modell, de USA eher das 2 und 3-Parameter-Modell). Das in Conquest verwendete 1-Parameter-Modell mit der Bestimmung der Plausible Values hat den großen Vorteil, dass es die besten Schätzungen für den Mittelwert und die Varianz eines Landes liefert. Wie schlecht Wuttke recherchiert hat, wird auch bei seinen Ausführungen „über den Programmierer Wilson“ deutlich. Adams habe auf den Programmierer in PISA 2003 keinen Zugriff mehr gehabt und musste sich am grünen Tisch zusammenreimen, was das Programm gemacht hat. Mark Wilson, der Wuttke offenbar nicht bekannt ist, ist Full Professor an der University of California, Berkeley, ein äußerst renommierter Professor, den man ohne Probleme kontaktieren kann, dessen Arbeitsgruppe Conquest weiter entwickelt und der für Ray Adams leicht erreichbar ist. Hätte Wuttke Einblicke in die Szene gehabt, wären ihm solche entlarvenden Äußerungen nicht passiert.

6.) Wuttke kritisiert weiter, dass die **Testdauer (2 Stunden) einen Effekt auf die Länderunterschiede** hatte, kürzere Testzeiten hätten auch andere Länderunterschiede zur Folge. Hier hat Wuttke Recht. Würde ein Bundesligist im Fußball gegen einen Landesligisten nur 10 Minuten spielen, würde das Spiel vermutlich nicht 13:0 sondern 3:0 ausgehen, vielleicht sogar 0:0. Die Frage wäre natürlich, ob man auf Grund der 10 Minuten die bessere Information über die Leistungsunterschiede erhalten hätte.

Zusammenfassung

Hätte sich Wuttke wirklich mit der einschlägigen Literatur und dem State of the Art in der Skalierung von großen Leistungstests auseinandergesetzt (Hunderte von Aufsätzen sind hier publiziert, von denen er kaum einen gelesen haben dürfte), wäre sein Urteil anders ausgefallen. Er hätte erkannt, dass das Vorgehen in PISA auf dem neuesten Stand der Diskussion ist und breiter wissenschaftlicher Konsens besteht, dass man es aktuell nicht besser machen kann.