Empirisches Arbeiten in der Deutschdidaktik: Qualifizierung für DoktorandInnen und PostdoktorandInnen
Modul 1, Jena, 4./5.2.2005

Protokoll Teil I:

Dr. Christof Nachtigall, Uni Jena, Lehrstuhl für Methodenlehre und Evaluationsforschung:
Quantitative Verfahren zur Erhebung von Schülerleistungen

In seiner **Einführung** problematisiert Nachtigall zunächst die gängige Trennung zwischen quantitativen und qualitativen Verfahren, die er als wenig nützlich charakterisiert. Er schlägt alternativ den Oberbegriff „Methoden der Datenerhebung / Messen" vor.
Er definiert „messen" als Verfahren

> „Relationen der erfahrbaren Welt (empirisches Relativ) derart in Zahlen (numerisches Relativ) abzubilden, dass die relevanten Relationen erhalten bleiben."

Es kommt also beim Messen darauf an, die Struktur, die vorliegt und die abgebildet werden soll, kategorial zu erfassen. Diese theoretische Arbeit muss vor jedem Messen erfolgen.
Die Messergebnisse sind selbst noch keine Quantitäten („Zahlen sind keine Zahlen"), sondern drücken immer nur Relationen aus. Ihnen muss vielmehr eine Qualität zugeschrieben werden.

Messniveaus (Skalenniveaus)

- kategorial → Nominalskala, z.B. Geschlecht
- ordinal → Ordinalskala, z.B. Schulnoten
- metrisch → Intervallskala, z.B. Körpergröße

Nachtigall verweist weiter auf die Bedeutung verschiedener Messniveaus (Skalenniveaus). Messeniveaus legen fest, welche Analysen mit den Daten sinnvoll bzw. sinnlos sind. Mittelwertbildungen etwa sind erst auf der Ebene metrischer Messung sinnvoll. Nachtigall betont auch hier die Bedeutung theoretischer Vorüberlegungen vor der eigentlichen Messung.

Wissenschaftliche Gütekriterien

- Objektivität → Messung ist unabhängig vom Messenden
- Reliabilität → Genauigkeit der Messung
  Reliabilität muss als Metaregel gefasst werden, weil es kein absolut reliables Instrument gibt, sondern es hier immer (nur) um die Perspektive der Optimierung gehen kann.
- Validität → wird gemessen, was gemessen werden soll? Auch das Kriterium der Vailidiät ist als Metaregel zu fassen. „Es gibt keine perfekt validen Tests."

Messmodelle

Die <u>klassische Testtheorie</u> geht von der „Formel":

> Gemessener Wert = wahrer Wert + Messfehler

Der „wahre Wert" bleibt dabei notwendig ein Konstrukt. Den Messfehler gilt es möglichst gering zu halten, was global bezogen auf den Gesamttest, nicht auf den Einzelfall möglich ist.

<u>Item Response Theorie (IRT)</u>

Hier liegt die „Formel" Lösungswahrscheinlichkeit = f(Fähigkeit, Schwierigkeit) zugrunde.

> „Die richtige Lösung einer Aufgabe (Item) ist probabilistisch abhängig von der Fähigkeit der Person und der Schwierigkeit der Aufgabe."

Die Konstruktion eines „wahren Wertes“ der klassischen Testtheorie gibt es hier nicht mehr. Dieses Verfahren ist aufwändiger, mathematisch anspruchsvoller und sinnvoll nur bei großen Stichproben (Faustregel: mehrere tausend Probanden).

Die Kompetenzen bei PISA sind auf der Basis dieser Testtheorie erstellt. Die Definition von einzelnen Stufen von Kompetenz erfolgt als Setzung „von außen“; einzelne Kompetenzstufen ergeben sich nicht aus der Messung selbst.
In der Diskussion wird auf die Verselbständigung der PISA Kompetenzstufen hingewiesen, in der vernachlässigt werde, dass es sich bei den Stufen um „Setzungen“ handelt.

5 Schritte der Testentwicklung am Beispiel des Thüringer Kompetenztests

Zieldefinition → was soll gemessen werden?
Textverstehen gemäß dem PISA-Modell
Operationalisierung durch Aufgabenpool → wie soll gemessen werden?
Aufgabenoptimierung → Eindeutigkeit der Fragestellung; Korrekturhinweise, Praktikabilität
Die Validität kann erhöht werden durch theoretische Überlegungen, die Einschätzung von Experten und die Berücksichtigung von Außenkriterien (z.B. Testbedingungen). Die Objektivität kann durch geschlossene Aufgabenformate sowie präzise Korrekturanweisungen erhöht werden. Bei kritischen Konstrukten werden mehrere Aufgaben bzw. Fragen zu einer Skala zusammengefasst, um die Reliabiltät zu erhöhen (Summenscoreberechnung)
Pilotierung (Vortest)
Ziel dieser Phase ist es, möglichst viele Problemstellen herauszufinden, indem der Test an einer erprobt wird und zusätzliche Fragebogen/Interviews mit den Probanden ausgewertet werden. Problematische Items sollen so herausgefunden werden.
Aufgabenselektion und Erstellung des endgültigen Tests
Hierzu gehört die Analyse der Aufgabenschwierigkeiten → sollten bei Tests, die diagnostisch eingesetzt werden sollen, streuen: d. h. wenige ganz leichte Aufgaben, viele Aufgaben mittleren Schwierigkeitsgrades, wenige schwere Aufgaben.
Analyse der Reliabilität (testen die Items, die ähnliches messen sollen, das auch tatsächlich?)

Testauswertung:
Im wesentlichen zwei Möglichkeiten: Deskriptive Analyse → Grafiken, Kennwerte bei Stichprobendaten
Schließende Statistik → Inferenzstatistik → gibt Signifikanzen an.

Vor allem drei Aspekte treten in der Diskussion immer wieder hervor:
Dies ist einmal die Frage, inwieweit Messinstrumente geeignet sind, die Qualität von Deutschunterricht tatsächlich zu verbessern. Die Frage verweist einmal auf die Güte der durchgeführten Testes selbst, zum anderen auf die Frage ob Tests Ergebnisse erbringen, die für den Deutschunterricht und seine konkreten schulischen Bedingungen hilfreich sind. Umgekehrt ist die Frage nach Wirkungen die Testkultur auf die Aufgabenkultur im Unterricht thematisch, etwa durch Trainings bestimmter Formate, die testüblich sind. Schließlich geht es um die Frage der Instrumentalisierung von Testergebnissen, wie sie in der Folge der PISA-Studie zu beobachten ist.