Willkommen zur Vorlesung Empirische Methoden I

5. Vorlesung: Vom Begriff zur Messung

Prof. Dr. Wolfgang Ludwig-Mayerhofer

Universität Siegen – Philosophische Fakultät, Seminar für Sozialwissenschaften

Begriff – Operationalisierung – Messung

- Zur Überprüfung von Hypothesen müssen die in ihnen verwendeten Begriffe empirisch ‚greifbar' gemacht werden. Dieser Vorgang heißt allgemein „Operationalisierung".
- Ziel der Datenerhebung ist letztlich, den beobachteten Phänomenen Symbole (meist: Zahlenwerte) zuzuordnen; diese liegen der Datenauswertung zu Grunde. Dieser Vorgang heißt Messung.
- Vor der Operationalisierung und Messung müssen Begriffe definiert und u. U. in ihrem dimensionalen Gehalt geklärt werde (Konzeptspezifikation)

In der qualitativen Forschung spielt dies alles keine Rolle! Konzepte entstehen dort ‚aus dem Datenmaterial'.

Operationalisierung als wichtige Aufgabe

Ein häufiger Anfänger-Fehler besteht darin, das Forschungsproblem nicht in eigene Operationalisierung zu ‚übersetzen'.

Beispiel: Leistungsdruck in der Schule

Hypothese (deskriptiv): In der Schule (konkret: Oberstufe) herrscht so hoher Leistungsdruck, dass den SchülerInnen nicht ausreichend Zeit für Freizeitaktivitäten bleibt

Nicht:	„Die Schule setzt mich häufig unter Leistungsdruck"
	„Die Schule behindert meine Freizeitgestaltung"
Sondern:	Schulanforderungen möglichst genau messen
	Freizeitverhalten und Freizeitwünsche messen

Konstrukt – Indikator

Häufig ist das Merkmal, welches wir erfassen wollen, nicht direkt beobachtbar. Wir sprechen hier von Konzept oder Konstrukt. Die empirische beobachtbaren Größen, anhand derer es erfasst werden soll, heißen Indikatoren.

Beispiel: Armut

Arm sind Personen, die „. . . über so geringe (materielle, kulturelle und soziale) Mittel verfügen, dass sie von der Lebensweise ausgeschlossen sind, die in dem Mitgliedsstaat, in dem sie leben, als Minimum annehmbar ist“ (Definition der EU)

Indikatoren: Einkommen? Ausstattung mit Konsumgütern? Lebenschancen (Bildung, Lebenserwartung)? Partizipation?

Konstrukt – Indikator: Weiteres Beispiel

Arbeitslosenquote: Anteil der „Arbeitslosen“ an allen Erwerbspersonen (Erwerbstätige plus Arbeitslose)

Zähler:

- Alle beim Arbeitsamt als arbeitslos gemeldeten Personen?
- Alle Arbeitssuchenden, die keinerlei Erwerbstätigkeit nachgehen?
- Arbeitslos Gemeldete und „Stille Reserve“?

Nenner:

- Alle Erwerbspersonen?
- Alle zivilen Erwerbspersonen?

Konzeptspezifikation

Zur Konzeptspezifikation gehört (a) die Definition (oder möglichst präzise Umschreibung) des jeweiligen Begriffs sowie (b) die Klärung möglicher (unterschiedlicher, aber miteinander verwandter) Dimensionen, die der Begriff enthält. Gelegentlich beschränken sich Autoren auch auf (b), weil dadurch der Gehalt eines Begriffs deutlich zu Tage tritt (siehe das zweite hier folgende Beispiel):

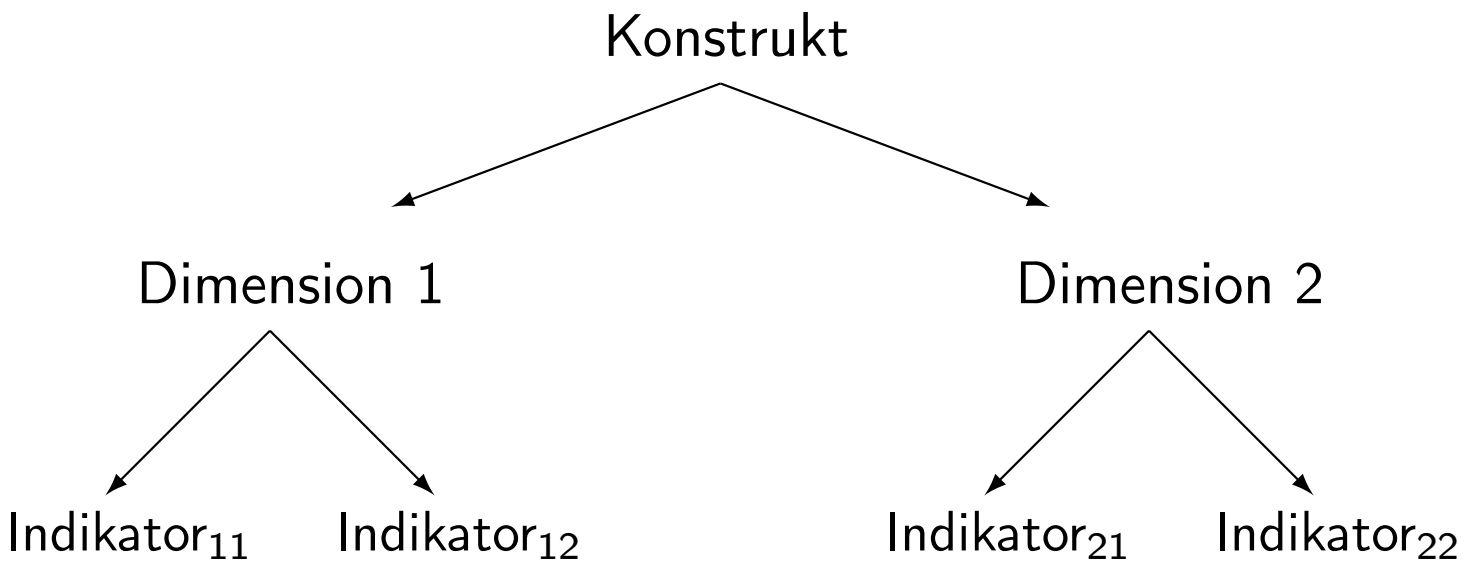
Beispiel: Armut (siehe oben): geringe materielle, kulturelle, soziale Ressourcen

Beispiel: Autoritarismus

- Neigung zu Unterwürfigkeit und Anpassung
- Feindseligkeit gegen Minderheiten und Schwächere
- Eintreten für „Ruhe und Ordnung“

Konzept – Dimension – Indikator

An die Konzeptspezifikation schließt sich die Operationalisierung an – die Angabe der spezifischen Indikatoren, mittels derer ein Konzept bzw. dessen einzelne Dimensionen gemessen werden sollen.



Messung mit mehreren Indikatoren

Es ist im allgemeinen empfehlenswert, Merkmale durch mehrere Indikatoren zu messen – jedenfalls wenn Messfehler zufällig sind. Solche Messungen führen (wenn korrekt durchgeführt) zu genaueren Messergebnissen.

Negativbeispiel: Prüfung (Klausur) mit nur einer einzigen Frage (aus breitem Stoff).

Mehrere Indikatoren müssen natürlich auch herangezogen werden, wenn das Konstrukt mehrdimensional ist (am besten: für jede Dimension wiederum mehrere Indikatoren)

Messung mit mehreren Indikatoren: Beispiel

Rechtsextremismus-Skala (Auszüge)

http://www.polwiss.fu-berlin/osz/dokumente/for_rechts.htm

- (Autoritarismus:) Wer seine Kinder zu anständigen Bürgern erziehen will, muss von ihnen vor allem Gehorsam und Disziplin verlangen.
- (Nationalismus:) Deutschland sollte wieder eine führende Rolle in der Welt übernehmen.
- (Ethnisch motivierte Fremdenfeindlichkeit:) Ausländer sollten so schnell wie möglich Deutschland verlassen.
- (Pronazistische Einstellungen:) Ohne Judenvernichtung würde man Hitler heute als großen Staatsmann ansehen.

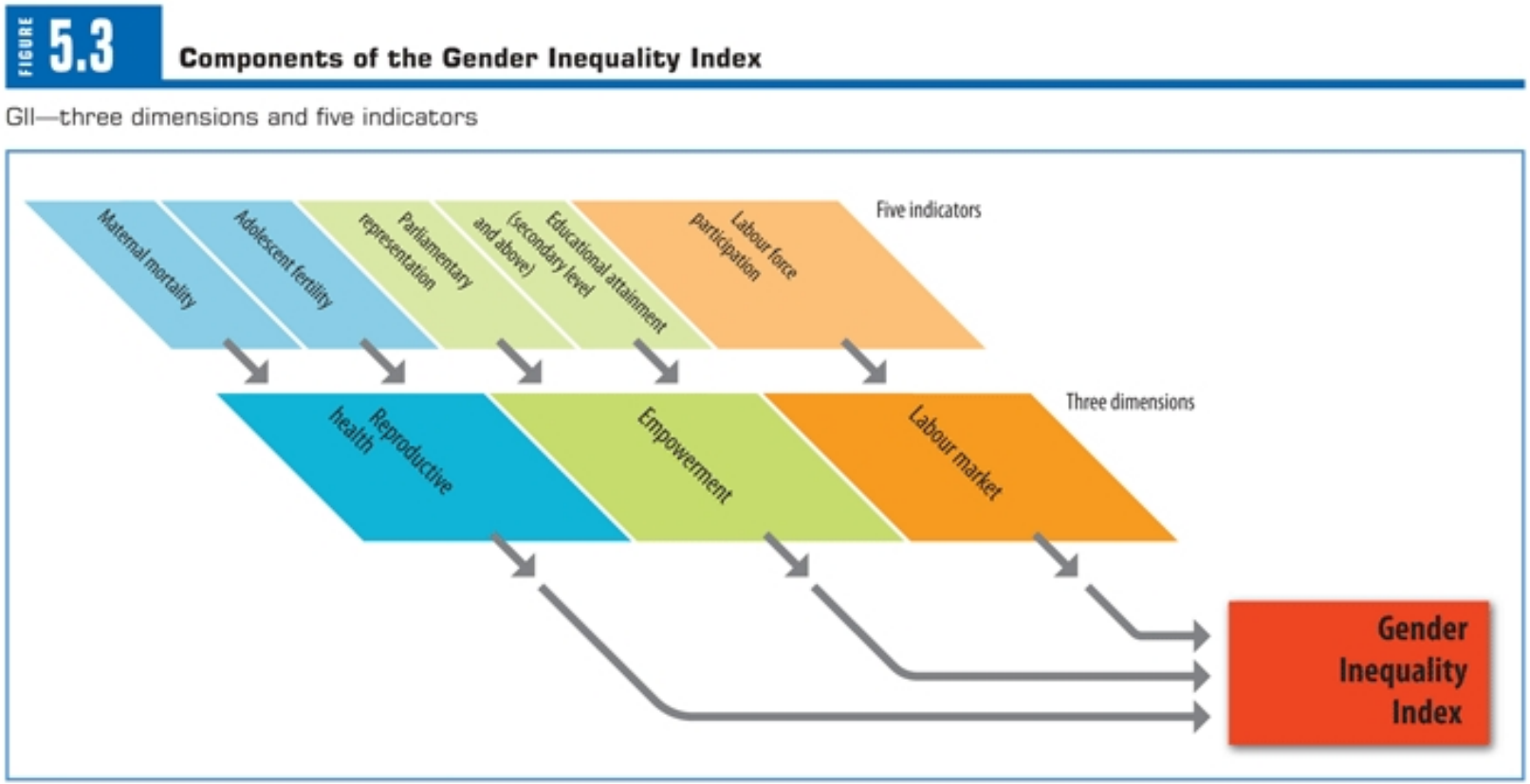
Messung mit mehreren Indikatoren: Index

Unter einem Index versteht man eine Messung anhand mehrerer Indikatoren, die jedoch manchmal relativ beliebig zusammengefügt werden. (Ein Beispiel auf den nächsten Seiten.)

Gelegentlich wird unter Index auch eine einfache Messgröße verstanden (z.B. Big-Mac-Index, siehe Diekmann).

Beispiel Index: Der GII

Zur Erfassung der Ungleichheit zwischen Frauen und Männern haben die UN (genauer: das United Nations Development Programm) den Gender Inequality Index entwickelt, siehe http://www.wunrn.com/news/2010/11_10/11_15_10/111510_gender.htm



Beispiel Index: Der GII

Aus dem Human Development Report 2010 der Vereinten Nationen, S. 156:

HDI rank	Gender Inequality Index[a] Rank	Gender Inequality Index[a] Value	Maternal mortality ratio[b]	Adolescent fertility rate[c]	Seats in parliament (%)	Population with at least secondary education (% ages 25 and older)		Labour force participation rate (%)		Contraceptive prevalence rate, any method (% of married women ages 15–49)	Antenatal coverage of at least one visit (%)	Births attended by skilled health personnel (%)
					Female	Female	Male	Female	Male			
	2008	2008	2003–2008[d]	1990–2008[d]	2008	2010	2010	2008	2008	1990–2008[d]	1990–2008[d]	2000–2008[d]
VERY HIGH HUMAN DEVELOPMENT												
1 Norway	**5**	**0.234**	7	8.6	36.1	99.3	99.1	77.3	82.6	88.4	..	..
2 Australia	**18**	**0.296**	4	14.9	29.7	95.1	97.2	69.9	83.0	70.8	..	99[e]
3 New Zealand	**25**	**0.320**	9	22.6	33.6	71.6	73.5	72.1	84.5	..	..	94[e]
4 United States	**37**	**0.400**	11	35.9	17.0[f]	95.3	94.5	68.7	80.6	72.8	..	99
5 Ireland	**29**	**0.344**	1	15.9	15.5	82.3	81.5	62.8	80.7	89.0	..	100
6 Liechtenstein	..	..	..	..	24.0	..	..	..	..	..	..	..
7 Netherlands	**1**	**0.174**	6	3.8	39.1	86.3	89.2	73.4	85.4	67.0	..	100
8 Canada	**16**	**0.289**	7	12.8	24.9	92.3	92.7	74.3	82.7	74.0	..	100
9 Sweden	**3**	**0.212**	3	7.7	47.0	87.9	87.1	77.1	81.8	..	..	..
10 Germany	**7**	**0.240**	4	7.7	31.1	91.3	92.8	70.8	82.3	..	..	100[g]
11 Japan	**12**	**0.273**	6	4.7	12.3	80.0	82.3	62.1	85.2	54.3	..	100
12 Korea, Republic of	**20**	**0.310**	14	5.5	13.7	79.4	91.7	54.5	75.6	80.2	..	100
13 Switzerland	**4**	**0.228**	5	5.5	27.2	62.9	74.5	76.6	87.8	..	..	100[g]
14 France	**11**	**0.260**	8	6.9	19.6	79.6	84.6	65.8	74.9	71.0	..	..
15 Israel	**28**	**0.332**	4	14.3	14.2	78.9	77.2	61.1	70.1	..	..	..

Messung mit mehreren Indikatoren: Skala

Unter einer Skala versteht man eine Messung anhand mehrerer Indikatoren, wobei spezifische Annahmen über die Struktur der Beobachtungen gemacht werden.

Diese Struktur sollte sich empirisch prüfen lassen.

Nachfolgend einige Beispiele hierzu.

Skalierungsverfahren I: Guttman-Skala

Eine Guttman-Skala besteht aus mehreren Items, wobei jedes Item eine „stärkere“ Ausprägung des jeweiligen Merkmals misst als das vorherige.

Beispiel: (Konventionelle) politische Beteiligung

Ich werde zur Wahl gehen	ja / nein
Ich werde Geld für den Wahlkampf spenden	ja / nein
Ich werde im Wahlkampf für eine Partei aktiv werben (Infostände, Verteilen von Materialien, usw.)	ja / nein
Ich werde für ein Mandat oder Amt kandidieren	ja / nein

Skalierungsverfahren II: Likert-Skala

Bei einer Likert-Skala werden mehrere Aussagen („Items“, „Statements“) vorgegeben; die Befragten geben auf einer mehrstufigen Antwortvorgabe (meist vier bis sieben Stufen) den Grad ihrer Zustimmung an.

Quelle: International Social Survey Programme, 2002

1. Wir möchten mit ein paar Fragen zur Berufstätigkeit von Frauen beginnen.

Inwieweit stimmen Sie den folgenden Aussagen zu oder nicht zu?

☞ ***Bitte in jeder Zeile ein Kästchen ankreuzen.***

	Stimme voll und ganz zu	**Stimme zu**	**Weder noch**	**Stimme nicht zu**	**Stimme überhaupt nicht zu**	Kann ich nicht sagen
Eine berufstätige Mutter kann ein genauso herzliches und vertrauensvolles Verhältnis zu ihren Kindern finden wie eine Mutter, die nicht berufstätig ist.	☐	☐	☐	☐	☐	☐
Ein Kind, das noch nicht zur Schule geht, wird wahrscheinlich darunter leiden, wenn seine Mutter berufstätig ist.	☐	☐	☐	☐	☐	☐
Alles in allem: Das Familienleben leidet darunter, wenn die Frau voll berufstätig ist.	☐	☐	☐	☐	☐	☐
Einen Beruf zu haben ist ja ganz schön, aber das, was die meisten Frauen wirklich wollen, sind ein Heim und Kinder.	☐	☐	☐	☐	☐	☐

Messniveaus (Skalenniveaus)

Achtung – der Begriff „Skala“ hat hier eine andere Bedeutung als auf den vorherigen Folien!

- Nominalskala: Unterschiedliche Ausprägungen bedeuten Unterschiedlichkeit – sonst nichts (Bsp.: Parteien).
- Ordinalskala: Unterschiedliche Ausprägungen bedeuten Reihenfolge – aber keine Angaben über Größe der Abstände (Bsp.: Olympiamedaillen, allgemein: Rangfolgen. Auch: Guttman-Skala, Likert-Skala [letztere oft wie intervallskaliert behandelt]).
- Intervallskala: Die Differenzen der Messwerte sind aussagekräftig (oder sollen es sein) – Bsp. Temperatur in °C.
- Ratioskala (Verhältnisskala): Die Verhältnisse der Messwerte sind aussagekräftig – Bsp. Temperatur in °K; evtl. Einkommen.

Gütekriterien I: Validität

Unter V. versteht man die Gültigkeit einer Messung – anders gesagt: dass gemessen wird, was gemessen werden soll.

Verfahren zur Prüfung:

- Inhaltsvalidität: Keine formale Prüfung – trotzdem wichtig
- Kriteriumsvalidität: Bezug auf Außenkriterium
  - Übereinstimmungsvalidität („Concurrent validity“): Außenkriterium wird gleichzeitig gemessen (Bsp.: Autoritarismusgrad bei Nazis)
  - Vorhersagevalidität: Außenkriterium wird später gemessen (Bsp.: Studienerfolg nach Aufnahmeprüfung)

Gütekriterien I, hier: Konstruktvalidität

Aufgrund vorliegender Theorien wird Zusammenhang zwischen verschiedenen Konstrukten angenommen und geprüft.

Bestätigung des angenommenen Zusammenhangs spricht für (nicht: beweist!) Validität.

Variante: Multi-Trait-Multi-Method Matrix

- Das gleiche Konstrukt, mit verschiedenen Methoden gemessen: Starker Zusammenhang (Konvergenz)
- Verschiedene Konstrukte: Schwacher oder kein Zusammenhang (Diskriminanz)
- Verschiedene Messungen des gleichen Konstruktes zeigen ähnliche Zusammenhänge mit anderen Merkmalen

Gütekriterien II: Reliabilität

Reliabilität heißt Zuverlässigkeit. Messinstrumente sollen bei wiederholter Messung am (unveränderten!) Objekt immer den gleichen Messwert liefern.

Formal: Zusammenhang zwischen beobachteten und wahren Werten. Da man wahre Werte nicht erheben kann, behilft man sich mit

- Test-Retest-Methode
- Paralleltest-Methode
- Prüfung interner Konsistenz

Gütekriterien III: Objektivität

Objektivität bezieht sich auf Messung durch Dritte und meint Unabhängigkeit der Messung von der Person, die sie durchführt.

Beispiele: Beurteilung von SchülerInnen durch LehrerInnen, PatientInnen durch ÄrztInnen, Medieninhalten durch CodiererInnen.

Genauer wird oft unterschieden zwischen:

- Durchführungsobjektivität: Die Person, die die Daten erhebt (z. B. Testleiter) hat keinen (bzw. immer den gleichen) Einfluss.
- Auswertungsobjektivität: Bestimmte Beobachtungen sollen gleich interpretiert werden, unabhängig von der Person des Beobachters.
- Interpretationsobjektivität: Die Schlüsse aus den Beobachtungen sollen unahängig von Person des Beobachters sein.

Gütekriterien in der qualitativen Forschung

In der qualitativen Forschung werden intensive Diskussionen geführt, welche Gütekriterien zur Anwendung kommen sollten und welche genaue Bedeutung diese in der qualitativen Forschung haben. Dabei werden sehr unterschiedliche Positionen eingenommen. Einig ist man sich nur, dass die Gütekriterien der standardisierten Forschung nicht eins zu eins übernommen werden können.

Ich beziehe mich im Folgenden nur auf Ausschnitte der Diskussion. Seien Sie nicht überrascht, wenn Sie in Lehrbüchern und anderswo ausführlichere Darstellungen mit zusätzlichen Punkten finden.

Validität in der qualitativen Forschung

Validität in der qualitativen Forschung kann nicht (wie in der standardisierten Forschung) die „Gültigkeit“ einzelner (punktueller) „Messungen“ (Daten) sein. Sie bezieht sich letztlich auf den gesamten Prozess der Forschung einschließlich der Interpretation der Daten (Daten sind nicht „an sich“ valide [oder nicht]).

Wenn man davon ausgeht, dass qualitative Forschung auf „Alltagskonstruktionen“ (oder „Konstrukte erster Ordnung“) abzielt, dann beruht Validität vor allem darauf, zeigen zu können, dass die Rekonstruktion dieser Konstrukte (die Erzeugung von „Konstrukten zweiter Ordnung“) den Alltagskonstruktionen angemessen ist; die „Standards alltäglicher Verständigung“ müssen klar sein.

Validität in der qualitativen Forschung: Verfahren

- Datenerhebungsmethoden: Müssen geeignet sein, Alltagskonstruktionen, Alltagspraktiken, Rituale usw. sichtbar werden zu lassen $\rightarrow$ offene Erhebungsmethoden der Beobachtung bzw. Befragung, die soziale Praxis bzw. deren Konstruktion durch Befragte ausreichend sichtbar werden lassen.
- Auswertung: Geschieht häufig in der Gruppe mit dem Ziel, unterschiedliche Perspektiven zu klären und zu konsensualer (und damit ‚richtiger' ?) Interpretation zu kommen (U. Oevermann: Objektive Hermeneutik; [zumindest phasenweise] Interpretation in Gruppen ist aber für nahezu alle Auswertungsverfahren sinnvoll).

Reliabilität in der qualitativen Forschung

PW-S:

- Rekonstruktion der alltäglichen Standards der Verständigung: Es muss (bspw.) *gezeigt* werden, dass Interviewer den Befragten Raum zur Entfaltung gegeben haben; unterschiedlichen Darstellungsformen muss durch Auswertung Rechnung getragen werden.
- Nachweis der „Reproduktionsgesetzlichkeit der Fallstruktur“ (innerhalb eines Falles oder über Fälle hinweg).

Uwe Flick (1995) betont vor allem die Kontrolle der Datenerhebung und -protokollierung. „Prozedurale Reliabilität“ kann gewährleistet werden bspw. durch

- gute Schulung der InterviewerInnen (siehe dazu Helfferich 2005)
- präzise Transkriptionsregeln

Triangulation

Triangulation: „. . . die Kombination verschiedener Methoden, verschiedener Forscher, Untersuchungsgruppen, lokaler und zeitlicher Settings sowie unterschiedlicher theoretischer Perpsektiven“ (Flick 1995, S. 330).

Nach Norman Denzin (1989):

- Daten-Triangulation: Daten an unterschiedlichen Orten, zu unterschiedlichen Zeitpunkten etc. erheben
- Forscher-Triangulation: Unterschiedliche Beobachter, Interviewer einsetzen
- Theorien-Triangulation: Verschiedene theoretische Perspektiven wählen
- Methoden-Triangulation: Verschiedene Methoden einsetzen.

Triangulation wird nicht nur als Verfahren der Steigerung von Validität gesehen, sondern als eines, die „Reichhaltigkeit“ der Ergebnisse (und damit ihre Qualität) zu steigern.

Literatur

Zu Gütekriterien in der qualitativen Forschung

Denzin, Norman (1989): The Research Act (3. Aufl.). Englewood Cliffs, NJ: Prentice Hall.

Flick, Uwe (1995): Qualitative Forschung. Theorie, Methoden, Anwendung in Psychologie und Sozialwissenschaften, Reinbek bei Hamburg: Rowohlt (oder neuere Auflage)

Helfferich, Cornelia (2005 [2. Auf.]): Die Qualität qualitativer Daten. Manual für die Durchführung qualitativer Interviews, Opladen: VS Verlag für Sozialwissenschaften