Zur Anwendung von State-Trait-Modellen in der wehrpsychologischen Eignungsdiagnostik unter besonderer Berücksichtigung computerunterstützter Tests

Rolf Steyer
Ivailo Partchev, Katrin Seiß, Sandra Menz und Thomas Hübner

Bericht Nr. 1 zum Forschungsvorhaben InSanI – 1190 – V 9000

Stand: 17. März. 2000

Anschrift der Autoren
Universität Jena
Institut für Psychologie
Lehrstuhl für Methodenlehre und Evaluationsforschung
Am Steiger 3, Haus 1
07743 Jena

Tel. 03641/945231
Fax 03641/945232
Email: Rolf.Steyer@uni-jena.de
Homepage: http://www.uni-jena.de/svw/metheval

Zur Anwendung von State-Trait-Modellen in der wehrpsychologischen Eignungsdiagnostik unter besonderer Berücksichtigung computerunterstützter Tests

Rolf Steyer, Ivailo Partchev, Katrin Seiß, Sandra Menz und Thomas Hübner
Institut für Psychologie, Universität Jena

Zusammenfassung

In diesem Bericht wird in einem einführenden *ersten Teil* begründet, warum die Latent-State-Trait-Theorie (LST-Theorie) auch für die wehrpsychologische Eignungsdiagnostik von Bedeutung ist. Das Kernargument dabei ist, dass auch in diesem Bereich der Diagnostik damit gerechnet werden muss, dass die Testergebnisse nicht nur messfehlerbehaftet sind, sondern auch Effekte der psycho-sozio-biologischen *Situation* beinhalten, in der die Probanden getestet werden, ebenso wie Effekte der *Interaktion* zwischen Person und Situation. Die aus einer Erhebung zu einer Messgelegenheit resultierenden Testwerte sind daher in aller Regel unzuverlässiger als die Fehlervarianzen (in der KTT) bzw. die Standardschätzfehler (in der IRT) widerspiegeln. Zur Fehlervarianz muss bei der KTT die Varianz der Situations-/Interaktionsvariablen hinzuaddiert werden und bei der IRT müssen die Standardschätzfehler ebenfalls um den Betrag erhöht werden, der durch die situativ bzw. interaktiv bedingte Variabilität der Fähigkeitszustände zustande kommt.

In einem *zweiten Teil* werden dann die Grundbegriffe der LST-Theorie referiert und ein Modell der LST-Theorie dargestellt, das erlaubt, die Varianzen und Varianzanteile der Traitvariablen, Statevariablen, Stateresiduen und Messfehler zu bestimmen. Diese Theorie und die referierten Modelle können dann angewandt werden, wenn zu jeder von mindestens zwei Messgelegenheiten mindestens zwei Messungen (z.B. zwei parallele Tests oder zwei Testhälften) des betrachteten Merkmals vorliegen.

In einem *dritten Teil* wird erstmals ein Modell der LST-Theorie konstruiert, das auch für Testwerte, die aus einem CAT-Verfahren resultieren, anwendbar ist. Dabei müssen zwar ebenfalls Messungen zu mindestens zwei Messgelegenheiten vorliegen, aber innerhalb jeder Messgelegenheit genügt ein einziger Testwert (eine einzige Messung), sofern der zugehörige Standardschätzfehler mitgeliefert wird. Das liegt daran, dass mit dem Stichprobenmittelwert der quadrierten Standardschätzfehler die (klassische) Fehlervarianz berechnet werden kann. Eine parallele Messung innerhalb derselben Messgelegenheit zur Ermittlung der Fehlervarianz (oder der Reliabilität) erübrigt sich daher.

In einem *vierten Teil* wird mit einer Simulationsstudie untersucht, bei welcher Stichprobengröße und Itemzahl das vorgestellte Verfahren hinreichend genaue Ergebnisse liefert. Dabei zeigt sich, dass die Varianz der Situations-/Interaktionsvariablen bei *großer Itemzahl* veridikal geschätzt werden kann, dass bei kleiner Itemzahl aber eine deutliche Unterschätzung festzustellen ist, deren Ursache in der Überschätzung der individuellen Standardschätzfehler liegt. Zu bedenken ist bei der Wertung dieser Ergebnisse allerdings, dass wir kein adaptives Verfahren, sondern die Vorgabe einer festen Itemmenge für alle Personen simuliert haben. Insbesondere ist also beim adaptiven Verfahren zu erwarten, dass die Itemzahl, bei der eine veridikale Schätzung der individuellen Standardschätzfehler zu erwarten ist, wesentlich geringer ist, als bei der hier gewählten fixierten Itemvorgabe.

Das vorgestellte Verfahren hat den Vorteil, dass es auch dann anwendbar ist, wenn jede Person im CAT-Verfahren unterschiedliche Items und sogar eine unterschiedliche Zahl von Items zu beantworten hat. Der Nachweis dieser Behauptung durch eine Simulationsstudie, die das CAT-Verfahren simuliert, steht allerdings aus.

Eignungsdiagnostik zielt auf die Erfassung der Eigenschaften von Personen, sogenannte *Traits*, die, neben anderen Ursachen, die nicht an die betrachtete Person gebunden sind (wie etwa situative Effekte), einen Einfluss auf das aktuelle Verhalten einer Person haben. Die derzeit in der Eignungsdiagnostik der Bundeswehr verwendeten Verfahren sind auf einem beachtenswerten Stand, indem sie weitgehend Papier und Bleistift durch den Computer als Erhebungsmedium ersetzt haben und indem sie von den modernen Entwicklungen der Psycho-

metrie, insbesondere dem adaptiven Testen auf der Basis von Modellen der Item-Response-Theorie, Gebrauch machen (Hornke, 1995; Storm, 1999; Sands et al., 1997).

In einer anderen Hinsicht besteht aber u. U. ein weiterer Entwicklungs- und Anpassungsbedarf, nämlich aus der Situationismus- und Interaktionismusdebatte in der Persönlichkeitspsychologie und den durch diese Debatte motivierten methodologischen Entwicklungen die methodischen Konsequenzen für die eigene Datenerhebung und –auswertung zu ziehen. Wie dies *möglich* ist, soll im vorliegenden Bericht aufgezeigt werden. Ob und wenn ja, in welchem Ausmaß dies *nötig* ist, gilt es im Projekt „Verbesserung der wehrpsychologischen Eignungsdiagnostik durch moderne mehrdimensionale Messmodelle“ (FV InSan I – 1099-V-9001) u. a. auch empirisch zu untersuchen.

Fassen wir die Ergebnisse der o. g. Debatte und die Vielzahl der damit verbundenen empirischen Befunde zusammen! Psychologische Diagnostik findet nicht im situationalen Vakuum statt. Alle psychologischen Test- und Messergebnisse sind in einem gewissen Maß *auch* von situativen Effekten und der Interaktion zwischen Person und der Situation, in der der Test erhoben wird, beeinflusst. Wie stark dies jeweils der Fall ist und wie praktisch bedeutsam dies ist, muss für jedes Testverfahren und jeden Anwendungszweck spezifisch untersucht werden. Für klassische Persönlichkeitstests wie das Freiburger Persönlichkeitsinventar (FPI), den NEO FFI oder das Eysenck Persönlichkeitsinventar gilt bspw., dass zwischen 5% und 10% der Varianz der Testwerte in den untersuchten Stichproben situativ und/oder interaktiv bedingt waren (Deinzer, Steyer, Eid, Notz, Schwenkmezger, Ostendorf und Neubauer, 1995). Ergebnisse zu Leistungstests liegen noch nicht vor, aber auch hier ist zu erwarten, dass ein nicht zu vernachlässigender systematischer Varianzanteil durch die Effekte von Situationen und der Interaktion zwischen Personen und Situationen bedingt ist. Diese Erwartung begründet sich u. a. auf die Alltagserfahrungen, dass Leistungen innerhalb der gleichen Person durchaus großen Schwankungen unterworfen sind. In dem Maße, in dem diese Erwartung zutrifft, ist die Genauigkeit der Schätzung der Personenkennwerte in den eignungsdiagnostischen Verfahren der Bundeswehr überschätzt. Die praktischen Konsequenzen wären für eine eventuelle Revision dieser Verfahren je nach Zweck der Testung und der Wichtigkeit richtiger diagnostischer Entscheidungen (z.B. Platzierung einfacher Soldaten vs. Auswahl von Piloten) zu überdenken.

Der vorliegende Bericht ist in drei Teile gegliedert. Im *ersten Teil* wird, ausgehend von der State-Trait-Unterscheidung in der Differentiellen Psychologie, die Latent-State-Trait-Theorie (LST-Theorie) dargestellt, eine methodologische Entwicklung in der Psychometrie, die durch die Situationismus- und Interaktionismusdebatte in der Persönlichkeitspsychologie motiviert ist. Diese Unterscheidung ist auch für die Eignungsdiagnostik der Bundeswehr relevant, da es dabei ja gerade darum geht, *Personeigenschaften* (Traits) zu ermitteln, die prädiktiv für aktuelles Verhalten im Alltag und im Einsatz sind.[1] Im *zweiten Teil* wird das

[1] Dieser erste Teil ist eine Umarbeitung einer Darstellung der LST-Theorie, die kürzlich in Reichle und Schmitt (1998) publiziert wurde. Eine weitere aktuelle Darstellung findet man in Steyer, Schmitt und Eid (1999).

CAT-Verfahren der Eignungsdiagnostik der Bundeswehr in den hier relevanten Aspekten dargestellt und es wird ein LST-Modell vorgestellt, das auf die aus dem CAT-Verfahren resultierenden Daten angewandt werden kann, sofern solche Daten aus mindestens zwei Messgelegenheiten vorliegen. Im *dritten Teil* schließlich demonstrieren wir mit einer Simulationsstudie, dass mit dem vorgeschlagenen Modell tatsächlich die situations- und interaktionsbedingte Varianz geschätzt werden kann.

1. Darstellung der Latent-State-Trait-Theorie

1.1 Die State-Trait-Unterscheidung

Herkömmlicherweise werden *States* und *Traits* unter Verwendung der Kriterien *Transsituationale Konsistenz* und *zeitliche Stabilität* unterschieden (z.B. Schmitt, 1990, S. 8). Diese Kriterien spannen einen zweidimensionalen Raum auf, in dem sich psychologische Variablen als Punkte darstellen lassen. So haben bspw. *Intelligenz* und *Neurotizismus* hohe Ausprägungen auf diesen beiden Dimensionen, wohingegen Stimmungen wie *Wohlbefinden* und *Innere Ruhe* vergleichsweise niedrige Ausprägungen darauf haben. Dennoch sind auf der einen Seite auch die Testwerte „typischer Traitvariablen“ (bspw. *Neurotizismus)* zu einem gewissen Ausmaß durch die Situationen mitdeterminiert, in denen die Tests erhoben werden (z.B. Deinzer et al., 1995). Auf der anderen Seite gibt es auch hinsichtlich „typischer Statevariablen“ (bspw. *Wohlbefinden*) starke interindividuelle Differenzen, die nicht situations-, sondern personbedingt sind (z.B. Eid, Notz, Steyer & Schwenkmezger, 1993; Majcen, Steyer & Schwenkmezger, 1988; Steyer, Schwenkmezger & Auer, 1990). Auch Untersuchungen anderer Konstrukte wie *Einstellungen* (Steyer & Schmitt, 1990a; Schmitt, Schwarz, Steyer & Schmitt, 1993), *Bewältigungsverhalten* (Ferring, Klauer, Filipp & Steyer, 1990) und *Soziale Erwünschtheit* (Schmitt & Steyer, 1993) sowie anderer psychologisch relevanter Merkmale wie *Cortisol im Speichel* (Kirschbaum, Steyer, Eid, Patalla, Hellhammer & Schwenkmezger, 1990) zeigen, dass für psychologisch relevante Observablen (beobachtbare Variablen) generell angenommen werden muss, dass sie durch mindestens zwei Faktoren determiniert werden: erstens Personen und zweitens Situationen und/oder Interaktionen zwischen Personen und Situationen.

Bei jeder State-Trait-Analyse gibt es mindestens zwei Ziele. Das *erste* ist die Information darüber, in welchem Ausmaß die betrachteten psychologischen Merkmale *traitlike* bzw. *statelike* sind. Das *zweite Ziel* besteht darin zu untersuchen, ob die verschiedenen Komponenten der betrachteten psychologischen Merkmale mit unterschiedlichen Variablen korreliert sind bzw. durch diese erklärt werden können. Die Traitkomponenten wären am ehesten durch andere Traitvariablen zu erklären oder könnten durch ihre Korrelation mit den Traitkomponenten anderer Konstrukte die Systematik gewisser Persönlichkeitsstrukturen belegen. Die Abweichungen der Statekomponenten von der Traitkomponente dagegen wären

am ehesten durch situationscharakterisierende Variablen zu erklären. Welche Merkmale einer Situation erklären die Abweichung des *aktuellen* (der Statekomponente) vom *habituellen* psychologischen Merkmal (der Traitkomponente)?

Forschungsprogrammatisch kann man State- und Traitkomponenten als *Grundkategorien psychologischer Theorien* ansehen. Generell sollten uns nicht die Observablen (i.d.R. sind das Testwertvariablen) selbst interessieren, sondern deren State- und deren Traitkomponenten. Eine Observable ist immer als eine Mischung verschiedener Komponenten anzusehen, deren Unterscheidung aber für eine sinnvolle psychologische Theorie und für die psychologische Diagnostik generell von Bedeutung ist. Sicherlich ist in der Regel bspw. die *Messfehlerkomponente* nur von technischem Interesse, inhaltlich-psychologisch gesehen aber irrelevant. Für manche Theorien sind eher die *Statekomponenten* bzw. die Abweichung der Statekomponente von der Traitkomponente, für andere dagegen eher die *Traitkomponenten* von Bedeutung. Die Mischung aus beidem, wie wir sie mit einer Observablen vorliegen haben, wird dagegen selten von zentralem Interesse sein.

Zur *Deskription* scheinen die eingangs genannten Kriterien *transsituationale Konsistenz* und *zeitliche Stabilität* zu taugen, aber zur *Definition* des Begriffs *Trait*? Impliziert nicht die transsituationale Konsistenz eine gewisse zeitliche Stabilität? Kann zeitliche Stabilität überhaupt als ein sinnvolles Kriterium zur *Definition* eines Traits verwendet werden? Hieße das nicht, ein Trait wäre unveränderbar, keine Reifung, kein Lernen, keine echten Traitveränderungen wären möglich? Würde man absolute Stabilität als definierendes Kriterium bejahen, dann bliebe in letzter Konsequenz nur übrig, einen Trait als eine bestimmte genetische Ausstattung, unabhängig von ihrer Manifestation im Verhalten und Empfinden, zu definieren. Dies wäre dann vielleicht ein sinnvoller *biologischer*, aber sicherlich kein *psychologischer* Begriff (Steyer, 1995).

1.2 Grundbegriffe der Latent-State-Trait-Theorie

In der Latent-State-Trait-Theorie (LST-Theorie) wird daher auf die Stabilität als *definierendes* Kriterium für einen Trait verzichtet. Lediglich die transsituationale Konsistenz ist für die Definition entscheidend. Relative zeitliche Stabilität ist zwar ein wichtiges deskriptives Merkmal, aber letztlich Konsequenz der transsituationalen Konsistenz und daher lediglich ein zutreffendes *Beschreibungsmerkmal*. Diese Stabilität kann über einen gewissen Zeitraum sogar perfekt sein, wenn bspw. kein für das betrachtete Merkmal relevantes Lernen oder Reifen stattfindet.

Der Traitbegriff der LST-Theorie setzt zunächst voraus, dass eine Observable (d.h. eine manifeste, direkt beobachtbare Variable), in der Regel eine Testwertvariable, sowie das Untersuchungssetting, in dem diese Variable erhoben wird, vorliegt. Formal kann man dann eine solche Observable als Zufallsvariable auffassen und man kann Messfehler, situative

Effekte und die Interaktion zwischen Person und Situation „konzeptuell ausfiltern“, um den Traitbegriff zu definieren. Wie lässt sich diese Idee konkretisieren?

In Tabelle 1 sind die wichtigsten Grundbegriffe der LST-Theorie zusammengestellt. Ausgangspunkt ist dabei der folgende Typ eines Zufallsexperiments: Es wird eine Person aus einer *Population* Ω_U *von Personen*[2] gezogen, an der dann zu n Messgelegenheiten Beobachtungen erhoben werden. Zu jeder dieser Messgelegenheiten realisiert sich eine Situation aus der betreffenden *Menge* Ω_{S_k} von *möglichen Situationen* und es wird eine Beobachtung aus einer *Menge* Ω_{O_k} *möglicher Beobachtungen* registriert. Dabei können sowohl die Situation als auch die Beobachtung mehrdimensional sein, d.h. beide Mengen, Ω_{S_k} und Ω_{O_k}, können ihrerseits Kartesische Produktmengen sein. So könnte z.B. die Situation simultan durch mehrere Aspekte definiert sein und die Beobachtung aus der Registrierung der Antworten auf mehrere Fragen in einem bestimmten Test bestehen, die auch zu mehreren Skalen aggregiert werden könnten, die sowohl gleiche (Parallelformen) als auch verschiedene Merkmale erheben sollen (vgl. Deinzer et al., 1995, S. 3).

Die *i*te *Observable*, die zur *k*ten Messgelegenheit erhoben wird, notieren wir mit X_{ik}. Solche Observablen sind meist durch die Auswertungsvorschriften eines Testverfahrens definiert. Daher nenne ich sie auch Testwertvariablen. Außer auf die Testwertvariablen X_{ik} können wir auch auf die *Projektionen* $U\colon \Omega \to \Omega_U$ und $S_k\colon \Omega \to \Omega_{S_k}$ zurückgreifen. Dabei gibt ein Wert $U(\omega)$ an, welche Person aus der Population Ω_U gezogen wurde und ein Wert $S_k(\omega)$, welche Situation sich zur *k*ten Messgelegenheit für die gezogene Person realisiert hat. Mit diesem begrifflichen Instrumentarium können wir nun die verschiedenen Grundbegriffe der LST-Theorie definieren.

Die *Latent-Statevariable* $\tau_{ik} := E(X_{ik} \mid U, S_k)$ ist die Regression (oder bedingte Erwartung) von X_{ik} auf U und S_k. Beide Faktoren, Person und Situation, sowie deren Interaktion determinieren also definitionsgemäß den Zustand einer Person bzgl. der *i*ten Observablen zur Messgelegenheit *k*. Der Statebegriff bzgl. einer Observablen resultiert also durch die Ausfilterung der *Messfehlervariablen* $\varepsilon_{ik} := X_{ik} - \tau_{ik}$.

Die *Latent-Traitvariable* $\xi_{ik} := E(X_{ik} \mid U)$ dagegen ist die Regression (oder bedingte Erwartung) von X_{ik} auf U. Definitionsgemäß determiniert nur der Person-Faktor den Trait einer Person bzgl. der *i*ten Observablen zur Messgelegenheit *k*. Der Traitbegriff bzgl. einer Observablen resultiert also durch die Ausfilterung der Messfehlervariablen und des *Stateresiduums* $\zeta_{ik} := \tau_{ik} - \xi_{ik}$ aus der Observablen X_{ik}. Man kann zeigen, dass das Stateresiduum ζ_{ik} aus situativen Effekten $E(X_{ik} \mid S_k)$ und aus der Interaktion zwischen Person und Situation $[E(X_{ik} \mid U, S_k) - E(X_{ik} \mid U) - E(X_{ik} \mid S_k)]$ besteht. Es gilt nämlich:

$$\zeta_{ik} := \tau_{ik} - \xi_{ik} = E(X_{ik} \mid U, S_k) - E(X_{ik} \mid U) \qquad (1.1)$$

[2] Formal gesehen ist das einfach eine Menge.

$$= E(X_{ik} \mid S_k) + [E(X_{ik} \mid U, S_k) - E(X_{ik} \mid U) - E(X_{ik} \mid S_k)]. \quad (1.2)$$

Eine Umordnung der Definitionsgleichungen führt zu den beiden *Dekompositionen der Variablen* $X_{ik} = \tau_{ik} + \varepsilon_{ik}$ und $\tau_{ik} = \xi_{ik} + \zeta_{ik}$ (s. Tab. 1). Demnach wird also eine Observable X_{ik} additiv in eine Latent-Statevariable τ_{ik} und eine Messfehlervariable ε_{ik} zerlegt und die Latent-

Tabelle 1. Grundbegriffe der Latent-State-Trait-Theorie

Die Menge der möglichen Ergebnisse des Zufallsexperiments

$\Omega = \Omega_U \times \Omega_{S_1} \times \ldots \times \Omega_{S_k} \times \ldots \times \Omega_{S_n} \times \Omega_{O_1} \times \ldots \times \Omega_{O_k} \times \ldots \times \Omega_{O_n}$

Observablen

$X_{ik}: \Omega \rightarrow \mathbb{R}$

Projektionen

$U: \Omega \rightarrow \Omega_U$	Personprojektion
$S_k: \Omega \rightarrow \Omega_{S_k}$	Situationsprojektionen

Latente Variablen

$\tau_{ik} := E(X_{ik} \mid U, S_k)$	Latent-Statevariable
$\varepsilon_{ik} := X_{ik} - \tau_{ik}$	Messfehlervariable
$\xi_{ik} := E(X_{ik} \mid U)$	Latent-Traitvariable
$\zeta_{ik} := \tau_{ik} - \xi_{ik}$	Latent-Stateresiduum

Dekomposition der Variablen

$X_{ik} = \tau_{ik} + \varepsilon_{ik}$

$\tau_{ik} = \xi_{ik} + \zeta_{ik}$

Dekomposition der Varianzen

$Var(X_{ik}) = Var(\tau_{ik}) + Var(\varepsilon_{ik})$

$Var(\tau_{ik}) = Var(\xi_{ik}) + Var(\zeta_{ik})$

Wichtige Kenngrößen

$Rel(X_{ik}) := Var(\tau_{ik}) / Var(X_{ik})$	Reliabilität
$Con(X_{ik}) := Var(\xi_{ik}) / Var(X_{ik})$	Konsistenz
$Spe(X_{ik}) := Var(\zeta_{ik}) / Var(X_{ik})$	Messgelegenheitsspezifität

Statevariable wird ihrerseits additiv in die Latent-Traitvariable ξ_{ik} und das Latent-State-residuum ζ_{ik} zerlegt.

Interessanterweise gehen in diese Zerlegungen keinerlei Annahmen ein, die sich in einer Anwendung als falsch erweisen könnten. Vielmehr handelt es sich um eine logische Konsequenz aus der *Definition* der vier latenten Variablen τ_{ik}, ε_{ik}, ξ_{ik} und ζ_{ik}. Noch interessanter ist, dass auch die entsprechende *Dekomposition der Varianzen* $Var(X_{ik}) = Var(\tau_{ik}) + Var(\varepsilon_{ik})$ und $Var(\tau_{ik}) = Var(\xi_{ik}) + Var(\zeta_{ik})$ auf keinerlei Annahmen beruht, sondern ebenfalls ausschließlich aus den obigen Definitionen und den allgemein gültigen Eigenschaften des Regressionsbegriffs (z.B. Steyer & Eid, 1993) abgeleitet werden kann.

Auf dieser allgemein gültigen Additivität der Varianzen beruhen die Definitionen der Reliabilität, Konsistenz und Messgelegenheitsspezifität (s. Tab. 1), die wichtigsten Kenngrößen der LST-Theorie. Die *Reliabilität* ist ein Maß für die Güte des durch die Observable X_{ik} repräsentierten Mess- bzw. Testverfahrens. Die *Konsistenz* beschreibt das Ausmaß, in dem die Observable X_{ik} durch den Personfaktor bedingt ist, wohingegen die *Messgelegenheitsspezifität* angibt, inwieweit Situation und Interaktion zwischen Person und Situation die Observable determinieren. Konsistenz und Messgelegenheitsspezifität addieren sich zur Reliabilität auf.

Abbildung 1 zeigt die in der LST-Theorie vorgenommenen Dekompositionen der Observablen X_{ik} (durch Vierecke dargestellt) in die verschiedenen LST-theoretischen Komponenten (dargestellt durch Kreise). Die Messfehlervariablen und die Latent-Stateresiduen sind nicht durch Kreise gekennzeichnet, zählen aber ebenfalls zu den LST-theoretischen, nicht direkt beobachtbaren Variablen (Gähde, Jagodzinski & Steyer, 1992).

Abbildung 1 macht deutlich, dass viel zu vielen theoretischen Variablen viel zu wenige Observablen gegenüberstehen. Ohne die Hinzufügung von Annahmen, die dieses Missverhältnis drastisch ändert, besteht keine Möglichkeit, aus den Verteilungen der Observablen (oder deren Kennwerten wie Mittelwerte, Varianzen, Korrelationen) etwas über die theoretischen Variablen (z.B. deren Mittelwerte, Varianzen und Korrelationen) ableiten zu können. Obwohl der Übersichtlichkeit halber nicht eingezeichnet, können alle Statevariablen, alle Stateresiduen, alle Traitvariablen und alle Messfehlervariablen in unbekannter Höhe jeweils untereinander korrelieren. Aus den Definitionen der LST-theoretischen Variablen τ_{ik}, ε_{ik}, ξ_{ik} und ζ_{ik} folgen jedoch für alle Indexpaare (i, k) und (j, l)

$$Cov(\varepsilon_{ik}, \tau_{jl}) = Cov(\varepsilon_{ik}, \xi_{jl}) = Cov(\varepsilon_{ik}, \zeta_{jl}) = Cov(\zeta_{ik}, \xi_{jl}) = 0. \qquad (1.3)$$

Demnach sind also die Messfehlervariablen mit den State- und Traitvariablen sowie mit den Latent-Stateresiduen unkorreliert. Außerdem sind auch die Latent-Stateresiduen mit den Latent-Traitvariablen unkorreliert. Dabei beachte man, dass es sich hier nicht um Annahmen handelt, die in irgendeiner Anwendung falsch sein könnten, sondern um logische Folgerungen

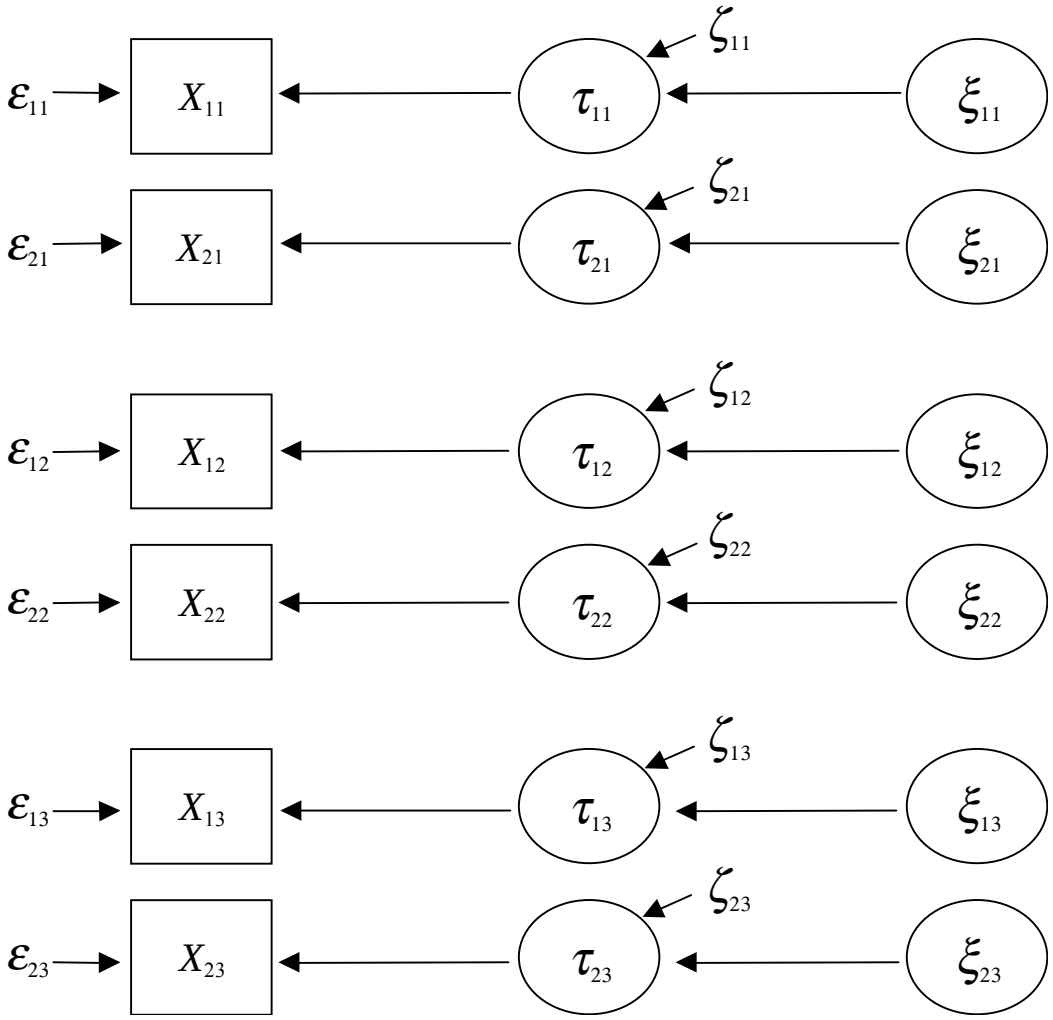

Abbildung 1. Die Dekomposition der Observablen X_{ik} in der Latent-State-Trait-Theorie

aus den Definitionen der LST-theoretischen Variablen. (Beweise hierzu, die auf den Eigenschaften bedingter Erwartungen beruhen, findet man z.B. in Steyer & Schmitt, 1990b).

1.3 Modelle der Latent-State-Trait-Theorie

Modelle der LST-Theorie entstehen durch die Einführung von Annahmen über die oben eingeführten Grundbegriffe. Repräsentieren bspw. die beiden Observablen X_{1k} und X_{2k} zu jeder Messgelegenheit k zwei parallele Tests, die den gleichen Zustand erheben sollen, so ist die Annahme

$$\tau_{1k} = \tau_{2k} =: \tau_k \qquad k = 1, ..., 3 \tag{1.4}$$

plausibel. Die gemeinsame latente Zustandsvariable zur Messgelegenheit k bezeichnen wir dann jeweils mit τ_k. Führen wir darüber hinaus noch die Annahme ein, dass die latenten

Eigenschaften sowohl zwischen den Observablen als auch über den betrachteten Messzeitraum hinweg identisch sind,

$$\xi_{1k} = \xi_{2k} = \xi_{1l} = \xi_{2l} =: \xi, \qquad k, l = 1, ..., 3 \tag{1.5}$$

so können wir die sechs Latent-Traitvariablen ξ_{ik} durch eine einzige Latent-Traitvariable ξ ersetzen. Weitere Vereinfachungen ergeben sich, wenn wir annehmen, dass die Situationen, in denen die States erhoben werden, zwischen den Messgelegenheiten unabhängig sind, oder dass wenigstens gilt:

$$Cov(\zeta_{ik}, \zeta_{jl}) = 0, \qquad k \neq l. \tag{1.6}$$

Eine letzte Vereinfachung ergibt sich, wenn die Messfehlervariablen ε_{ik} als unkorreliert angenommen werden können:

$$Cov(\varepsilon_{ik}, \varepsilon_{jl}) = 0, \qquad (i, k) \neq (j, l). \tag{1.7}$$

In Abbildung 2 ist das Pfaddiagramm dieses Modells angegeben. In verschiedenen Arbeiten (z.B. Steyer, Ferring & Schmitt, 1992; Eid, 1995; Eid, Steyer & Schwenkmezger, 1996) wurden alternative, weniger restriktive Annahmen dargestellt, die ebenfalls zu Modellen führen, die erlauben, die LST-theoretischen Kenngrößen zu bestimmen und das resultierende Modell zu testen, indem man die Implikationen für die Struktur der Kovarianzmatrix der Observablen auf Übereinstimmung mit deren empirischer Kovarianzmatrix untersucht.

Darüber hinaus gibt es natürlich auch Modelle (z.B. Eid & Hoffmann, 1998), in denen Latent-Traitvariablen vorkommen, die sich über die Zeit verändern. Allerdings braucht man für solche Modelle mindestens zwei relativ weit auseinander liegende Erhebungsphasen, innerhalb derer mindestens zweimal eine Latent-Statevariable mit mindestens je zwei (Parallel-)Tests erhoben wird. Die beiden Erhebungsphasen sollten deshalb relativ weit auseinander liegen, damit eine wahre Traitveränderung auch möglich ist. Was „relativ weit" heißt, hängt vom betrachteten Konstrukt und den sonstigen Bedingungen ab, insbesondere davon, wie schnell dabei eine tatsächliche Traitveränderung erwartet werden kann. In bestimmten Phasen kann eine Entwicklung (d.h. eine wahre Traitveränderung) mehr oder auch weniger schnell vonstatten gehen.

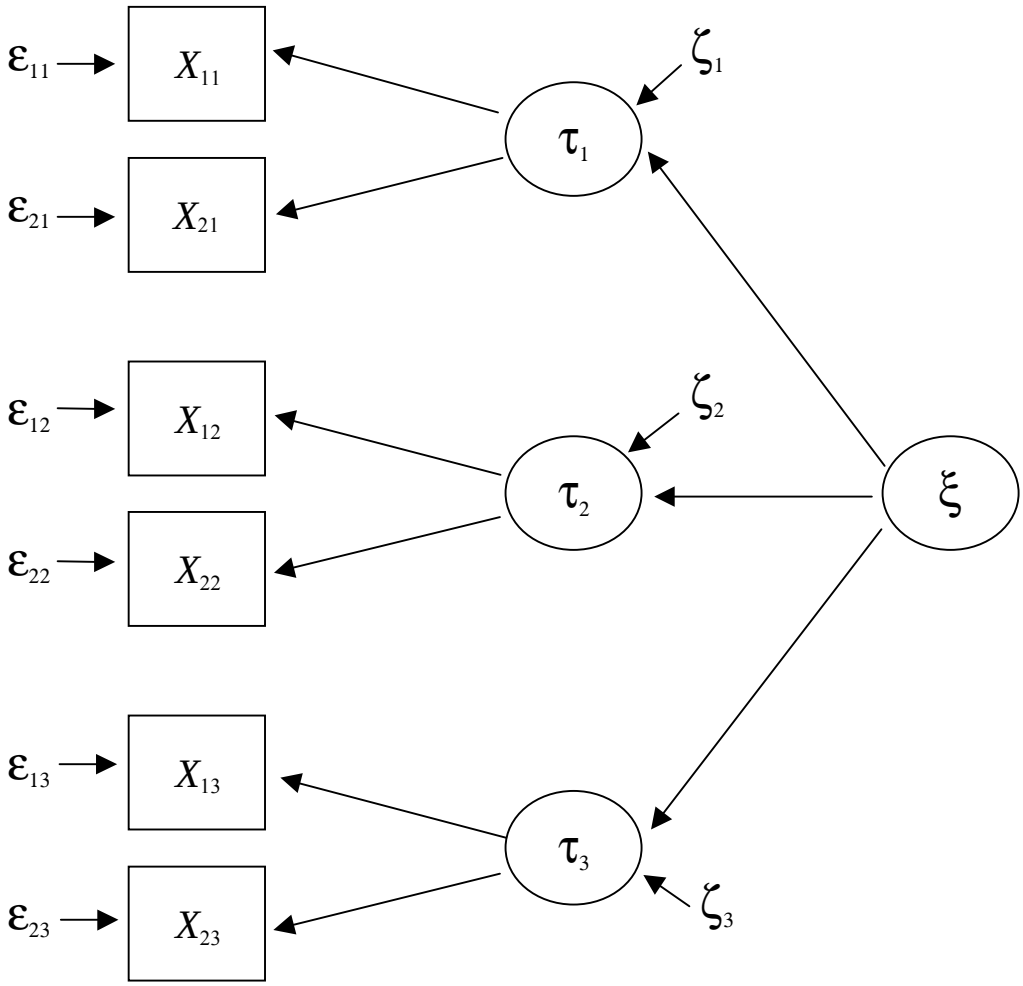

Abbildung 2. Das Singletrait-Multitstate-Modell mit zwei Observablen pro Messgelegenheit

In diesem Kontext sind auch Modelle von großem Interesse, in denen wahre Traitveränderungsvariablen oder wahre Stateveränderungsvariablen als latente Variablen vorkommen und direkt mit anderen Variablen erklärt werden können (Steyer, Eid & Schwenkmezger, 1997; Steyer, Partchev & Shanahan, 2000). Die Grundidee solcher Modelle ist sehr einfach. Betrachten wir ein Modell, in dem die Observablen X_{ik} zu jeder Messgelegenheit k in eine gemeinsame State- und eine jeweils spezifische Messfehlerkomponente zerlegt werden, bei dem also gilt:

$$X_{ik} = \tau_k + \varepsilon_{ik}, \quad i = 1, ..., p, \quad k = 1, ..., n. \tag{1.8}$$

Dabei ist leicht zu erkennen, dass auch die folgenden Gleichungen gelten müssen:

$$X_{i1} = \tau_1 + \varepsilon_{i1}, \tag{1.9}$$

$$X_{i2} = \tau_2 + \varepsilon_{i2} = \tau_1 + (\tau_2 - \tau_1) + \varepsilon_{i2}, \tag{1.10}$$

$$X_{i3} = \tau_3 + \varepsilon_{i3} = \tau_1 + (\tau_2 - \tau_1) + (\tau_3 - \tau_2) + \varepsilon_{i3}, \text{ etc.} \tag{1.11}$$

Übersetzt man diese Gleichungen in ein entsprechendes Pfaddiagramm, erkennt man sofort, dass wir nun latente Stateveränderungsvariablen (in den obigen Gleichungen mit Klammern markiert) als latente Variablen in das Modell eingeführt haben. Deren Abhängigkeit von anderen Variablen kann dann direkt in einem entsprechenden Strukturgleichungsmodell untersucht werden.

Ganz entsprechend lässt sich diese Grundidee auch auf wahre Traitveränderungen übertragen. Betrachten wir bspw. ein Modell, in dem die Latent-Statevariablen τ_k innerhalb jeder von zwei Erhebungsphasen in eine gemeinsame Trait- und eine jeweils spezifische Messfehlerkomponente zerlegt werden:

$$\tau_k = \xi_1 + \zeta_k, \quad k = 1, 2 \quad \text{und} \quad \tau_k = \xi_2 + \zeta_k, \quad k = 3, 4. \tag{1.12}$$

Dann ist leicht zu erkennen, dass auch die folgenden Gleichungen gelten müssen:

$$\tau_k = \xi_1 + \zeta_k, \quad k = 1, 2 \tag{1.13}$$

$$\tau_k = \xi_1 + (\xi_2 - \xi_1) + \zeta_k, \quad k = 3, 4. \tag{1.14}$$

Auch hier erkennt man sofort, dass wir nun eine latente Traitveränderungsvariable (in der obigen Gleichung mit Klammern markiert) als latente Variable in das Modell eingeführt haben. Auch deren Abhängigkeit von anderen Variablen kann man direkt in einem entsprechenden Strukturgleichungsmodell untersuchen. Die Verallgemeinerung auf mehr als zwei Erhebungsphasen geschieht in völliger Analogie zum Fall der Stateveränderung.

Unproblematisch ist auch eine Verallgemeinerung für kongenerische Messmodelle, in denen die Observablen unterschiedliche Ladungen auf den State- und Traitvariablen haben dürfen. Allerdings ist dabei eine Konstanz der Ladungen über die Zeit bei gleichen Messinstrumenten erforderlich. Einzelheiten hierzu findet man in Steyer et al. (2000).

1.4 Konsequenzen aus der LST- Theorie für die Individualdiagnostik

Modelle der LST-Theorie ermöglichen nicht nur die Schätzung des Varianzanteils, der durch den Trait erklärt wird, sondern berücksichtigen auch den Einfluss messgelegenheitsspezifischer, situationaler und/oder interaktionaler Effekte. Um diese Modelle anwenden zu können, bedarf es wiederholter Messungen desselben Konstruktes zu jeder von mindestens zwei Messgelegenheiten. Wenn davon ausgegangen wird, dass für jede Person eine natürliche Variation der Situationen zwischen den Messgelegenheiten besteht, ist es möglich, den Einfluss der Situationen und/oder Interaktionen zwischen Personen und Situationen zu schätzen. (Dies geschieht in ähnlicher Weise wie die Schätzung der Messfehlervarianz, bei der verschiedene Messungen desselben Konstruktes innerhalb einer Messgelegenheit vorgenommen werden.) Im Gegensatz zu anderen Ansätzen, die auf varianzanalytischen Modellen basieren, bedarf es im Rahmen der LST-Theorie keines Wissens über die Situationen, in

denen die Messung stattfindet. Aus diesem Grund stellen LST-Modelle eine wichtige Ergänzung zu varianzanalytischen Methoden dar.

Natürlich auftretende Situationen und/oder Interaktionen zwischen Personen und Situationen haben einen Einfluss auf die Messung von Traits (s. z.B. Deinzer et al., 1995). Dies ist von beträchtlichem praktischen Interesse: Wenn wir das Konfidenzintervall für den wahren Traitwert einer Person schätzen wollen, sollte der Konsistenzkoeffizient anstelle des Reliabiliätskoeffizienten genutzt werden. Dies ist eine ganz praktische Konsequenz, die auch für die Vorgehensweisen bei der Eignungsdiagnostik der Bundeswehr zu ziehen ist. Darüber hinaus kann es sinnvoll sein, Daten verschiedener Messgelegenheiten zu aggregieren, wenn der Einfluss der Situationen auf die Messungen reduziert werden soll (dies wurde von Epstein (1977; 1979; 1980; 1983; 1984; 1986; Epstein & O`Brien, 1985) vorgeschlagen und in den mathematischen Konsequenzen von Steyer und Schmidt (1990) beschrieben). Schließlich sollte immer bedacht werden, dass wir nicht von *der* Reliabilität, *der* Konsistenz und *der* Messgelegenheitsspezifität eines Messinstruments sprechen können, denn diese Koeffizienten können zwischen Messgelegenheiten und zwischen Studien variieren. Wenn bspw. Daten einerseits im Kontext eines psychologischen Experiments und andererseits in einer natürlich auftretenden Situation erhoben werden, werden wir unterschiedliche LST-Koeffizienten schätzen. Deshalb stellt sich die Frage, ob wir die LST-Koeffizienten, die wir im situationalen Kontext eines Experiments schätzen, zu Beurteilungen in einem anderen situationalen Kontext heranziehen können. Die Antwort auf diese Frage lautet: „Nein, das können wir nicht!“. Ebenso wie wir aus den Koeffizienten, die an einer Menge von Personen geschätzt wurden, nicht auf die Koeffizienten einer vollständig anderen Menge von Personen schließen können, können wir nicht aus einer Menge von Situationen auf eine andere Menge von Situationen schließen. Folglich sollte bspw. bei der Standardisierung eines psychologischen Testverfahrens nicht nur die Population der untersuchten Personen repräsentativ für die in Zukunft mit diesem Messinstrument zu untersuchenden Personen sein, sondern auch die Menge der *Situationen* muss repräsentativ für die Menge der Situationen zukünftiger Anwendungen sein. Dies ist nicht der Fall, wenn ein Test in einer Laborsituation standardisiert wird und dann auf Personen angewandt wird, die sich bspw. für eine Position in einem Unternehmen bewerben.
Hieraus folgt, dass LST-Koeffizienten psychologischer Messinstrumente in jenen situationalen Kontexten untersucht werden sollten, in denen das Instrument in Zukunft angewendet werden wird. Ebenso wie es nützlich erscheint, einen Test für verschiedene Subpopulationen von Personen wie bspw. Altersgruppen zu normieren, können verschiedene Normen auch für unterschiedliche situationale Kontexte wie bspw. Bewerbungssituationen, psychiatrische Behandlungen u.s.w. hilfreich sein.

Zusammenfassend kann Folgendes festgehalten werden: Die LST-Theorie liefert einen theoretischen Hintergrund und methodische Vorgehensweisen zur Schätzung von Koeffizienten für Reliabilität, Konsistenz und Messgelegenheitsspezifität für Messinstrumente in korrelativen, nichtexperimentellen Studien. Die Theorie verdeutlicht, dass die Population der

Situationen ein definierender Bestandteil des Zufallsexperiments ist, in dem psychologische Traits gemessen werden. Es ist davon auszugehen, dass jede Messung eines Traits von situationalen und/oder interaktionalen Effekten beeinflusst wird, denn Personen können niemals außerhalb von Situationen untersucht werden. Diese Tatsache muss bei Reliabilitäts- und Validitätsstudien Berücksichtigung finden.

2. Ein LST-Modell zur Analyse von CAT-Daten

2.1. Das CAT-Schätzverfahren

Das computerunterstützte adaptive Testverfahren (CAT-Schätzverfahren) der Bundeswehr liefert für jede Person einen Testwert und den zugehörigen Standardschätzfehler. Der Testwert ist dabei nicht etwa nur die Summe der richtigen Antworten, sondern wird aufgrund eines Modells der Item-Response-Theorie geschätzt (Hambleton & Swaminathan, 1985). Die für die weiteren Erörterungen wichtige Besonderheit des CAT-Verfahrens ist, dass jede Person eine andere Teilmenge der Items einer Itembank zu lösen hat und dass selbst die Anzahl der Items, die bearbeitet werden müssen, zwischen den Personen unterschiedlich ist (s. z.B. Storm, 1999; Wainer et al., 1990).

Dies rührt daher, dass einer Person so viele Items zur Bearbeitung vorgegeben werden, bis der Standardfehler der Schätzung des Personenparameters einen festgelegten Wert unterschreitet. Dieser Wert wird von unterschiedlichen Personen eben mit unterschiedlichen Itemanzahlen unterschritten. Unterschiedliche Items werden den Personen deswegen vorgelegt, damit sie optimal der Fähigkeit der Person angepasst sind. Eine sehr fähige Person sehr leichte Items bearbeiten zu lassen würde nur wenig zusätzliche Information über ihre tatsächliche Fähigkeitsausprägung liefern, genauso wie es sehr wenig Information bringen würde, einer Person mit sehr niedriger Fähigkeit ein sehr schwieriges Item vorzulegen. Diesen Zusammenhang beschreibt die Informationsfunktion des betreffenden Modells der Item-Response-Theorie (s. z.B. Steyer & Eid, 1992, S. 224; Hambleton & Swaminathan, 1985, pp. 98-99; Baker, 1992, 72-83).

Storm (1999) beschreibt das CAT-Verfahren wie folgt:

> Zu Beginn einer adaptiven Testung wird zunächst ein mittelschweres Item mit dem Wert von $b \approx 0$ vorgegeben und die Fähigkeit wird zunächst bei $\theta \approx 0$ angenommen. Löst der Pb diese Aufgabe, erhält er vorläufig eine um einen zunächst festen Betrag erhöhte Fähigkeitsschätzung ($\theta = +0{,}25$), sonst eine etwas erniedrigte ($\theta = -0{,}25$). Aufgrund dieser angenommenen Fähigkeitsschätzungen werden Aufgaben mit vergleichbaren Schwierigkeitsparametern aus der Itembank gezogen. Diese Vergabe fester Beträge wird dabei nur solange wiederholt, bis sogenannte Beantwortungsvarianz auftritt, d.h., wenn sich eine Beantwortungsabfolge „richtig/falsch“ ergibt. Im Folgenden werden dann keine festen θ-Beträge mehr vergeben, da der adaptive Algorithmus nun mittels eines mathematischen Vorganges mit der Schätzung der Fähigkeit beginnt. Vereinfacht ausgedrückt wird dabei aus der gegebenen Beantwortungssequenz (z.B. richtig/richtig/falsch/etc.) hinsichtlich der Aufgaben, deren Schwierigkeitsparameter ja bekannt sind (z.B. $b = -0{,}17$ / $b = +0{,}26$ / $b = +0{,}78$ / etc.), die plausibelste Fähigkeitsschätzung ermittelt. So resultiert z.B. in diesem Fall nach zwei richtig und einer falsch beantworteten Matrizenaufgabe eine Fähigkeitsschätzung von $\theta \approx 0{,}7$. Demnach erhält der Pb in dieser Phase der

Testung als nächstes eine Aufgabe mit der Schwierigkeit von $b \approx 0{,}7$. Würde der Pb solche Aufgaben im Folgenden mal richtig mal falsch beantworten, dann würden ihm entsprechend seinem Fähigkeitsniveau Aufgaben in einem Schwierigkeitsbereich von $b \approx 0{,}4$ bis $b \approx 1{,}0$ vorgegeben werden.

2.2 Das CAT-Schätzverfahren aus der Sicht der Klassischen Testtheorie (KTT)

Was ist die Ausgangssituation? Wie oben dargelegt, erhält man als Ergebnis der Testung eines Probanden zu einer Messgelegenheit sowohl eine *Schätzung* (d.h. einen Wert des Schätzers $\hat{\theta}_u$) *des Personenparameters* θ_u als auch eine Schätzung des *Standardschätzfehlers* $\sigma_{\hat{\theta}_u}$. Gehen wir der Einfachheit halber davon aus, dass $\hat{\theta}_u$ ein erwartungstreuer Schätzer des Personenparameters θ_u ist,[3] dann können wir – wie in der Klassischen Testtheorie – diesen Schätzer in seinen wahren Wert θ_u und die (personenspezifische) Messfehlervariable ε_u zerlegen:

$$\hat{\theta}_u = \theta_u + \varepsilon_u \tag{1.15}$$

Beim Standardschätzfehler $\sigma_{\hat{\theta}_u}$ handelt es sich um nichts anderes als um die Standardabweichung $\sigma_{\hat{\theta}_u} = \sigma_{\varepsilon_u}$ der (intra-individuellen) Verteilung des Schätzers $\hat{\theta}_u$. Das Quadrat des Standardschätzfehlers ist die *Person-bedingte Fehlervarianz* $Var(\varepsilon_u) = \sigma^2_{\varepsilon_u}$ des Schätzers $\hat{\theta}_u$ des Personenparameters.

Betrachtet man diesen Sachverhalt im Rahmen des für die KTT typischen Zufallsexperiments „Ziehe eine Person aus der Population und erhebe deren Leistung“, so entfällt der Personenindex *u* und man kann die aus der klassischen Testtheorie bekannte Gleichung

$$\hat{\theta} = \theta + \varepsilon \tag{1.16}$$

schreiben, wobei $\hat{\theta}$ der (manifesten) Testwertvariablen entspricht, θ der True-score-Variablen und ε der Messfehlervariablen. Für diese neue Messfehlervariable ε gilt: $Var(\varepsilon \mid U{=}u) = Var(\varepsilon_u) = \sigma^2_{\varepsilon_u} = \sigma^2_{\hat{\theta}_u}$. Bildet man nun den Erwartungswert $E[Var(\varepsilon \mid U)]$ über die Verteilung der Personen, so erhält man damit die *unbedingte Fehlervarianz* $Var(\varepsilon) = \sigma^2_{\varepsilon}$, wie wir sie aus der Klassischen Testtheorie kennen. Daraus kann man z.B. die *Reliabilität* der manifesten Testwertvariablen $\hat{\theta}$ berechnen:

$$Rel_{\hat{\theta}} = \frac{\sigma^2_{\theta}}{\sigma^2_{\hat{\theta}}} \tag{1.17}$$

(s. Rost, 1996, S. 354 oder auch Steyer & Eid, 1993, S. 231).

Fassen wir zusammen: Das CAT-Verfahren liefert die Schätzungen der Personenparameter und die zugehörigen Standardschätzfehler. Aus der Sicht der KTT sind die Schät-

[3] Dass die Schätzer des Personenparameters nicht erwartungstreu sind, ist bekannt (s. z.B. Rost, 1996, S. 304ff; Hoijtink & Boomsma, 1966). Hier gehen wir jedoch aus didaktischen und pragmatischen Gründen zunächst von dieser Annahme aus.

zungen der Personenparameter Werte einer fehlerbehafteten Testwertvariablen $\hat{\theta}$, deren Reliabilität man betrachten kann.[4] Der Erwartungswert der quadrierten Standardschätzfehler ist die Fehlervarianz, wie wir sie aus der KTT kennen.

2.3 Das CAT-Schätzverfahren aus der Sicht der Latent-State-Trait-Theorie (LST-Theorie)

Aus der Sicht der LST-Theorie sind die Werte der oben eingeführten Testwertvariablen $\hat{\theta}$, d.h. die Personparameter θ_u, bei einem einmalig durchgeführten Test nur ein *situativ verfälschter Traitparameter* der betreffenden Person. Aus der Perspektive der LST-Theorie charakterisiert der Parameter θ_u nicht die getestete Person, sondern die getestete *Person in der Situation*, in der die Testung stattfand.

Bei der wiederholten Testung eines Probanden erhält man wiederum eine Schätzung seines Personenparameters und des zugehörigen Standardschätzfehlers. Wird eine ganze Stichprobe von Probanden wiederholt getestet, hat man also sowohl von jedem Probanden zu jedem der beiden Messzeitpunkte jeweils die Schätzung des Personenparameters und seines Standardschätzfehlers, als auch auf Stichprobenebene die unbedingte Fehlervarianz und die Reliabilität zu jedem der beiden Messzeitpunkte. Darüber hinaus kann man aus diesen Daten die (Stichproben-)Kovarianz und die Korrelation zwischen den Schätzungen des Personenparameters zum ersten und zum zweiten Messzeitpunkt berechnen.

Wir bezeichnen nun mit $\hat{\theta}_k$, $k = 1, 2$, die beiden Schätzungen des Personenparameters zur Messgelegenheit k. Der Index i entfällt hier, da wir davon ausgehen, dass das CAT-Verfahren nur einen einzigen Testwert pro Messgelegenheit liefert. Unter den Annahmen des Singletrait-Multistate-Modells

$$\hat{\theta}_1 = \xi + \zeta_1 + \varepsilon_1, \tag{1.18}$$

$$\hat{\theta}_2 = \xi + \zeta_2 + \varepsilon_2, \tag{1.19}$$

$$Cov(\varepsilon_1, \varepsilon_2) = 0, \tag{1.20}$$

$$Cov(\zeta_1, \zeta_2) = 0 \tag{1.21}$$

(s. Abb. 3) ergeben sich folgende Gleichungen für die Kovarianz der beiden $\hat{\theta}$-Variablen:

---

[4] Die Testwertvariable $\hat{\theta}$ ist sicherlich den vielfach verwendeten Summenwerten vorzuziehen, die sich bestenfalls mit pragmatischen Argumenten rechtfertigen lassen. Der schönere Anwendungsfall der KTT ist der, mit dem CAT-Verfahren einen modellbasierten Testwert zu errechnen und mit der Reliabilität das Ausmaß seiner Fehlerbehaftetheit *mit einer einzigen Zahl* anzugeben. Der aus dem CAT-Verfahren resultierende Standardschätzfehler, die Wurzel aus der personenbedingten Fehlervarianz, ist natürlich viel informativer, da er für die jeweilige Person spezifisch ist. Der Nachteil ist jedoch gerade diese Spezifität. Bei 1000 Pbn hat man 1000 Standardschätzfehler. Insofern ergänzen sich das Reliabilitätskonzept der KTT und der Standardschätzfehler der IRT. Zugunsten des CAT-Verfahrens ist allerdings auch noch zu bemerken, dass dabei ja ein Maximum für den Standardschätzfehler festgelegt wird und damit ebenfalls eine wichtige Information über die Zuverlässigkeit der Messungen in einer einzigen Zahl enthalten ist.

$$Cov(\hat{\theta}_1, \hat{\theta}_2) = Cov(\xi + \zeta_1 + \varepsilon_1, \xi + \zeta_2 + \varepsilon_2) = Var(\xi). \quad (1.22)$$

Die Varianzen der beiden Stateresiduen ζ_1 und ζ_2 ergeben sich dann aus:

$$Var(\zeta_k) = Var(\hat{\theta}_k) - Var(\xi) - Var(\varepsilon_k), \quad k = 1, 2. \quad (1.23)$$

Schätzungen für alle Parameter auf der rechten Seite dieser Gleichung sind bekannt bzw. können den obigen Überlegungen zufolge schon bei zweimaliger Testung einer Stichprobe mit adaptiven Tests berechnet werden. Damit sind bei einem solchen Datenerhebungsverfahren alle Parameter des Singletrait-Multistate-Modells schätzbar, einschließlich der Konsistenz, der Messgelegenheitsspezifität und der Reliabilität.

Mit dieser Information lässt sich schon *grob beurteilen*, mit welcher Unzuverlässigkeit man bei der Schätzung der Traitparameter der Probanden rechnen muss, die additiv zur messfehlerbedingten Unzuverlässigkeit hinzukommt. Aus Sicht der LST-Theorie sind ja die Schätzwerte des Personparameters aus den Ergebnissen eines einmalig durchgeführten Tests nur Schätzungen des *situativ verfälschten Traitparameters* der betreffenden Person, nämlich von $\theta_k = \xi + \zeta_k$. In der Begrifflichkeit der LST-Theorie ist die situativ verfälschte Latent-Traitvariable die *Latent-State-Variable*. Diese kann man als fehlerbehaftete Schätzung für die Latent-Traitvariable ansehen. Situative Effekte und interaktionale Effekte zwischen Personen und Situationen, d.h. die Werte von ζ_k, werden im diesem Kontext der Schätzung eines Traitwerts also als Messfehler angesehen. Das Ausmaß der Fehlerbehaftetheit der Latent-State-variablen ist durch die Varianz von ζ_k charakterisiert. Fasst man nun die Schätzwerte des Personparameters aus den Ergebnissen eines einmalig durchgeführten Tests als Schätzwerte des Latent-Trait-Werts des betreffenden Probanden auf, so besteht die Fehlerkomponente aus der Summe $\zeta_k + \varepsilon_k$ des Latent-Stateresiduums ζ_k und der eigentlichen Messfehlervariablen ε_k. Eine grobe Abschätzung des Standardschätzfehlers bei der Schätzung des Latent-Traitwerts eines Probanden u ist demnach die positive Wurzel aus der Varianz der Summe $\zeta_k + \varepsilon_k$. Da diese beiden Variablen unkorreliert sind, gilt $Var(\zeta_k + \varepsilon_k) = Var(\zeta_k) + Var(\varepsilon_k)$, und für den Standardschätzfehler:

$$SE(\hat{\theta}_{ku}) \cong \sqrt{Var(\zeta_k) + Var(\varepsilon_k)}\,. \quad (1.24)$$

Dabei beachte man, dass man hier deswegen nur eine *grobe* Abschätzung der Unzuverlässigkeit der Schätzung des Traitwerts des Probanden u hat, weil dabei interindividuelle Unterschiede in der Genauigkeit der Schätzung der (situativ verfälschten) Traitwerte nicht berücksichtigt werden. Allerdings dürfte der Fehler nicht allzu groß sein, da für den reinen Messfehlervarianzanteil $Var(\varepsilon_k)$ im CAT-Verfahren ein Maximum festgelegt wird, dass für alle Personen gleich ist. Wie stark oder gering die Werte der personenbedingten Situations- und Interaktionsvarianz $Var(\zeta_k | U)$ um ihren Erwartungswert $Var(\zeta_k)$ herum variieren, bleibt bei diesem Verfahren unbeantwortet.

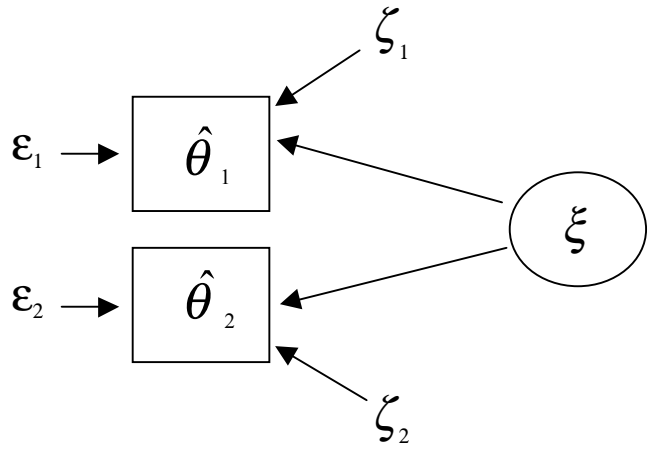

Abbildung 3. Das Singletrait-Multitstate-Modell mit einer Observablen pro Messgelegenheit

3. Eine Simulationsstudie

3.1 Das der Datenerzeugung zugrundeliegende Modell

Zur Illustration des oben dargestellten Verfahrens zur Bestimmung der situativen und/oder interaktiven Varianz im CAT-Verfahren haben wir eine Simulationsstudie durchgeführt. Die Daten wurden dabei nach der folgenden Modellgleichung erzeugt:

$$P(Y_{ik} = y_{ik} \,|U, S_k) = \frac{exp\,[y_{ik}(\xi + \zeta_k - \kappa_i)]}{1 + exp\,(\xi + \zeta_k - \kappa_i)} \tag{1.25}$$

Dabei bezeichnet $P(Y_{ik} = y_{ik}|U, S_k)$ die bedingte Wahrscheinlichkeit für eine Lösung ($Y_{ik} = 1$) bzw. Nichtlösung ($Y_{ik} = 0$) des Items i zur Messgelegenheit k als Funktion der Personvariablen U und der Situationsvariablen S_k. Bei dieser Gleichung handelt es sich um das klassische Rasch-Modell, das allerdings nun zu mehreren Messgelegenheiten k betrachtet wird und in dem die (verfälschte) Traitvariable θ_k durch die Summe der Traitkomponente ξ und der Situations- und Interaktionsvariable ζ_k ersetzt wurde. Die Parameter κ_i bezeichnen jeweils die Schwierigkeit des betreffenden Items i. Das Modell (1.25) stellt eine Erweiterung des dichotomen Rasch-Modells innerhalb der Item-Response-Theorie dar, bei dem die Lösungswahrscheinlichkeit für ein Item nicht nur gegeben eine Person, sondern gegeben eine *Person in einer Situation* betrachtet wird. Auch bei gegebener Person gehen wir dabei davon aus, dass sich zu einer gegebenen Messgelegenheit k verschiedene Situationen realisieren können, die dann unterschiedliche Werte der Variablen ζ_k zur Folge hätten.

3.2 Datenerzeugung

Es werden 10000 Fälle (Personen) zu $k = 3$ Messgelegenheiten betrachtet. Zu jeder Messgelegenheit k werden jeder Person 43 verschiedene Items vorgegeben. Alle Personen erhalten dieselben Items.[5] Die Lösungswahrscheinlichkeiten dieser Items lassen sich nach Gleichung (1.25) berechnen.

In der Simulationsstudie werden die theoretischen Parameter [s. die rechte Seite von (1.25)] dieses Modells aufgrund eines Algorithmus vorgegeben, den wir im Folgenden beschreiben werden. Anhand der so vorgegebenen theoretischen Größen, die für jede Person verschieden sind, wird das Antwortverhalten der Testpersonen simuliert [s. die rechte Seite von (1.25)]. Mit diesen künstlich erzeugten Daten wird dann so verfahren, wie es bei der Analyse empirischer Daten üblich ist. Unter Voraussetzung der Gültigkeit des Modells werden die theoretischen Parameter aus den Daten mittels eines Programms zur Analyse von Rasch-Modellen geschätzt.[6] Wenn diese aus den (künstlich erzeugten) Daten geschätzten Parameter mit den wahren Parametern übereinstimmen, die durch den Algorithmus vorgegeben wurden, können wir von der Richtigkeit des vorgeschlagenen Verfahrens ausgehen.

Der Algorithmus kann wie folgt beschrieben werden:

1. Zunächst werden die Schwierigkeiten von 43 (virtuellen) Items pro Messgelegenheit festgelegt. Dabei wählten wir den Bereich von −3 bis +3 mit gleichen Abständen zwischen den Schwierigkeiten. Ordnen wir die Items nach ihren Schwierigkeiten, dann hat also das erste Item die Schwierigkeit −3, das zweite die Schwierigkeit −2.85714, das dritte Item die Schwierigkeit −2.71429, das vierte Item die Schwierigkeit −2.57143 usw. Das letzte und 43te Item hat die Schwierigkeit +3. Mit dieser Wahl der Schwierigkeiten ist der gesamte Fähigkeitsbereich insofern abgedeckt, als dass es extrem unwahrscheinlich ist, dass eine Person alle oder kein einziges Item löst (s. den nächsten Punkt).
2. Als nächstes wird aus einer standardnormalverteilten Zufallsgröße die „wahre“ Fähigkeit, d.h. die Werte von ξ, für alle Testpersonen zufällig gezogen. Dabei resultieren also 10000 normalverteilte Werte mit Erwartungswert 0 und Varianz 1.
3. Bei der Erzeugung der Werte der Situations- und Interaktionsvariablen ζ_k nehmen wir wiederum deren Normalverteilung mit Erwartungswert 0 an. Weiter gehen wir davon aus, dass sich die Testpersonen in ihrer Variabilität bzw. Instabilität, d.h. in ihren bedingten Varianzen $Var(\zeta_k | U = u)$, unterscheiden. Wir gehen also davon aus, dass die personenbedingten Varianzen, also die Werte von $Var(\zeta_k | U)$, für verschiedene Personen durchaus unterschiedlich sein können, d.h. es soll nur $E[Var(\zeta_k | U)] = Var(\zeta_k)$, nicht aber $Var(\zeta_k | U) = Var(\zeta_k)$ gelten. Zwischen verschiedenen Messzeitpunkten dagegen sollen bei gegebener Person *keine* unterschiedlichen Varianzen der Situations-/Interaktionsvariablen ζ_k vorkom-

[5] Dies entspricht zwar nicht dem Sachverhalt, wie wir ihn im CAT-Verfahren vorliegen haben, erleichtert aber die Studie erheblich, da nicht schon im Datengenerierungsverfahren ein adaptives Schätzverfahren verwendet werden muss. Eine solche aufwendigere Simulationsstudie soll im Projekt in der zweiten Hälfte des Jahres 2000 durchgeführt werden.

[6] Dabei haben wir WINMIRA (von Davier, 1997a, b) verwendet.

men: $Var(\zeta_k|U) = Var(\zeta|U)$, $k = 1, 2, 3$.[7] In einem zweiten Schritt wird die Varianz von $Var(\zeta|U = u)$ für jede Person aus einer χ^2-verteilten Zufallsgröße mit einem Freiheitsgrad gezogen. Dabei gehen wir davon aus, dass ξ und $Var(\zeta_k|U)$ unabhängig voneinander sind, d.h. es besteht kein Zusammenhang zwischen der Ausprägung der Fähigkeit einer Person und der Instabilität bzw. Variabilität dieser Fähigkeit über die verschiedenen Messgelegenheiten.

4. In einem nächsten Schritt wird aus diesen zufällig ermittelten Werten die positive Quadratwurzel gezogen. Diese wird mit der (willkürlich festgelegten) Zahl 0.3 multipliziert. Diese Zahlen werden als Standardabweichung $Std(\zeta|U = u)$ genommen. Deren Quadrate haben die Eigenschaft: $E[Var(\zeta|U)] = Var(\zeta) = 0.30$. Da die Situations-/Interaktionsvariablen ζ_k voraussetzungsgemäß normalverteilt sind und den Erwartungswert 0 haben, liegt mit $Std(\zeta|U = u)$ eine personspezifische Normalverteilung vor. Für jede Person werden nun aus dieser Normalverteilung drei Werte gezogen, die drei Realisierungen von ζ_1, ζ_2 und ζ_3 für die jeweils betrachtete Person.
5. Ein weiterer Schritt besteht nun darin, dass für jede Person ihre „wahre“ Fähigkeit (ihr ξ-Wert) pro Messgelegenheit mit dem jeweiligen ζ_k-Wert addiert wird, woraus sich für die betreffende Person der situationsbedingt verfälschte wahre Fähigkeitswert zur Messgelegenheit k ergibt. Dabei handelt es sich um die Werte von θ_k.
6. Dann werden zu jeder der drei Messgelegenheiten 43 Items unterschiedlicher Schwierigkeiten κ_i vorgegeben. Diese Schwierigkeiten sind um null zentriert (s. oben Punkt 1).
7. Aus der Itemcharakteristik (1.25) wird die zu jedem der 3×43 Items gehörige Lösungswahrscheinlichkeit berechnet. Aus diesen Lösungswahrscheinlichkeiten wird ein manifestes Antwortverhalten generiert, indem man mittels eines Zufallszahlengenerators mit der betreffenden Lösungswahrscheinlichkeit die Zahl 1 und der Gegenwahrscheinlichkeit die Zahl 0 erzeugt. Auf diese Weise entstehen für jede Person eine Folge von $3 \times 43 = 129$ Nullen und Einsen, insgesamt also eine Datenmatrix vom Typ 10000×129.

Als Ergebnis der Datenerzeugung stehen von jeder der 10000 Personen neben der gerade beschriebenen Datenmatrix auch die folgenden theoretischen Parameter zur Verfügung:

- Der wahre Wert einer Person bzgl. der Fähigkeit ξ
- die drei Werte von ζ_1, ζ_2 und ζ_3 und damit die drei Werte von θ_1, θ_2 und θ_3, da $\theta_k = \xi + \zeta_k$
- die personbedingte Varianz $Var(\zeta_k|U = u) = Var(\zeta|U = u)$ und damit die personbedingte Streuung $Std(\zeta|U = u)$.

Die wesentlichen Parameter, die es in der statistischen Analyse zu reproduzieren gilt, sind zum einen die Varianz von ξ, die gleich 1 ist, und zum anderen die Varianz von ζ, die gleich 0.30 gesetzt wurde.[8]

---

[7] Die Werte von $Var(\zeta|U)$ kann man als Instabilität bzw. Variabilität der Person hinsichtlich der betrachteten Fähigkeit interpretieren.

[8] Schön wäre es natürlich auch, wenn die personenbedingten Varianzen $Var(\zeta|U = u)$ reproduziert werden könnten. Dies scheint uns aber bei nur zwei oder drei Messgelegenheiten ein aussichtsloses Unterfangen. Diese

3.3 Datenanalyse

Mittels WINMIRA wurden jeweils Teile der oben beschriebenen Datenmatrix analysiert. Dabei resultieren für jede Messgelegenheit pro Person ein Schätzwert von θ_1, θ_2 und θ_3 mit den zugehörigen Standardschätzfehlern. Diese sind dann die Datenbasis für die oben beschriebene State-Trait-Analyse, die uns die Schätzungen für die Varianz von ξ und die Varianz der ζ_k liefern soll. Bei der WINMIRA-Analyse haben wir ein zweifaktorielles Design gewählt. Der erste Faktor ist die *Itemzahl*, d.h. die Anzahl der Items, die pro Messgelegenheit für die Schätzung von θ_k verwendet wurde. Für diesen Faktor wählten wir die Ausprägungen 7, 15, 22 und 43 Items. Der zweite Faktor war die *Stichprobengröße*, wofür wir die Ausprägungen 1000, 2000, 5000 und 10000 wählten. Dabei sollte untersucht werden, bei welcher Stichprobengröße und welcher Itemzahl wir eine zuverlässige Schätzung der Varianz von ζ_k erhalten. Die Ergebnisse sind in Tabelle 2 zusammengestellt.

Für die Schätzung der *Varianz der Traitvariablen* ξ wurde die folgenden Schätzformel verwendet:

$$\hat{Var}(\xi) = \frac{1}{3}\left[\hat{Cov}(\hat{\theta}_1,\hat{\theta}_2) + \hat{Cov}(\hat{\theta}_1,\hat{\theta}_3) + \hat{Cov}(\hat{\theta}_2,\hat{\theta}_3)\right]. \tag{1.26}$$

Dabei bezeichnet $\hat{Cov}(\hat{\theta}_k,\hat{\theta}_l)$ die Stichprobenkovarianz der Schätzer der (situativ/interaktiv verfälschten) Personenparameter zur Messgelegenheit *k* bzw. *l*. Dabei beachte man, dass die *theoretischen* Kovarianzen $Cov(\hat{\theta}_k,\hat{\theta}_l)$ alle drei gleich $Var(\xi)$ sind. Ganz analog erfolgt die Schätzung $\hat{Var}(\hat{\theta})$ der Varianz der Schätzer $\hat{\theta}_k$ nach:

$$\hat{Var}(\hat{\theta}) = \frac{1}{3}\left[\hat{Var}(\hat{\theta}_1) + \hat{Var}(\hat{\theta}_2) + \hat{Var}(\hat{\theta}_3)\right]. \tag{1.27}$$

Auch die theoretischen Varianzen $Var(\hat{\theta}_k)$ sind alle drei gleich, was auch hier die einfache Mittelung rechtfertigt. Die Schätzung der Fehlervarianz erfolgt durch

$$\hat{Var}(\varepsilon) = \frac{1}{3}\left[\frac{1}{n}\sum_{u=1}^{n}\hat{SE}(\hat{\theta}_{1u})^2 + \frac{1}{n}\sum_{u=1}^{n}\hat{SE}(\hat{\theta}_{2u})^2 + \frac{1}{n}\sum_{u=1}^{n}\hat{SE}(\hat{\theta}_{3u})^2\right]. \tag{1.28}$$

Auch hier sind die theoretischen Fehlervarianzen für alle drei Messgelegenheiten gleich. Schließlich bleibt noch die Schätzung $\hat{Var}(\zeta)$ der Varianz von ζ, die nach der Formel

$$\hat{Var}(\zeta) = \hat{Var}(\hat{\theta}) - \hat{Var}(\xi) - \hat{Var}(\varepsilon) \tag{1.29}$$

erfolgte.

Einschätzung würde sich ändern, wenn wir viele (vielleicht 10 oder mehr) Messgelegenheiten vorliegen hätten. Unter inhaltlichen Gesichtspunkten wäre eine solche personenbedingte Varianz durchaus informativ, da sie ja als Variabilität oder Instabilität der betrachteten Fähigkeit der betreffenden Person interpretiert werden kann. Die (unbedingte) $Var(\zeta)$ dagegen gibt nur Auskunft über die durchschnittliche Variabilität der Fähigkeit in der betrachteten Population von Personen.

Während die Gleichungen (1.26) bis (1.29) ausschließlich auf Stichprobendaten beruhen, wie wir sie auch im normalen empirischen Fall vorliegen haben, können wir weitere Schätzungen betrachten, wenn wir auch die Information mitverwenden, die wir nur deswegen haben, weil die Daten simuliert sind. Dazu gehören z.B. die situationsbedingt verfälschten wahren Fähigkeitswerte für jeden der drei Messzeitpunkte, d.h. pro Person die Werte der drei Variablen θ_k (s. Punkt 5 im Abschnitt 3.2 Datenerzeugung). Insbesondere können wir mit dieser Information die Fehlervarianz der Schätzer $\hat{\theta}_k$ schätzen, und zwar über die Formel:

$$\widehat{Var}_{sim}(\varepsilon) = \frac{1}{3}\left[\frac{1}{n}\sum_{u=1}^{n}(\hat{\theta}_{1u} - \theta_{1u})^2 + \frac{1}{n}\sum_{u=1}^{n}(\hat{\theta}_{2u} - \theta_{2u})^2 + \frac{1}{n}\sum_{u=1}^{n}(\hat{\theta}_{3u} - \theta_{3u})^2\right]. \qquad (1.30)$$

Diese Formel gibt die richtige Information über die tatsächliche Fehlervarianz der Schätzer $\hat{\theta}_k$, zumindest bei sehr großen Personenstichproben. Sie ist allerdings eben nur in Simulationsstudien anwendbar. Formel (1.28) dagegen ist auch bei normalen Datenanalysen anwendbar, wobei aber zu befürchten ist, dass sie nur bei relativ großer Itemzahl zu veridikalen Schätzungen der Fehlervarianz führt. Die den Schätzungen der Standardschätzfehler zugrundeliegenden Theoreme enthalten nur Aussagen über die Asymptotik (s. z.B. Steyer & Eid, 1993, S. 231 oder Fischer, 1974, S. 297; Hoijtink & Boomsma, 1996), insbesondere nur: „für Zahl der Items gegen unendlich ...“. Die Ergebnisse der Simulationsstudie bestätigen diese Befürchtung (s. unten).

Betrachten wir Tabelle 2, kommen wir zu folgenden Beobachtungen: Die Güte der Schätzung der Varianz von ξ hängt offenbar nur von der Stichprobengröße ab. Schon bei einer Itemzahl von 7 liegt die Schätzung schon sehr nahe an der theoretischen Varianz von 1. Die Güte der Schätzung scheint nicht mit der Itemzahl zu wachsen.

Dies ist ganz anders bei der Schätzung $\widehat{Var}(\zeta)$ der Varianz von ζ. Hier ist offenbar nicht die Stichprobengröße, sondern die Anzahl der verwendeten Items für die Güte der Schätzung verantwortlich. Die besten Schätzungen werden erzielt, wenn wir alle 43 Items verwenden, und zwar unabhängig von der Größe der Personenstichprobe. Bei kleiner Itemzahl fällt die Schätzung $\widehat{Var}(\zeta)$ offenbar systematisch zu niedrig aus.

Die Größe der Fehlervarianzschätzung $\widehat{Var}(\varepsilon)$ hängt erwartungsgemäß von der Zahl der Items ab. Die Schätzung der Fehlervarianz von 1.2 bei 10000 Personen und sieben Items entspricht einer Reliabilität von .486 [s. Gl. (1.17)]. Wichtiger aber ist, dass die Schätzung der Fehlervarianz $\widehat{Var}(\varepsilon)$ erst bei großer Itemzahl zuverlässig und veridikal wird. Bei kleiner Itemzahl dagegen fällt die Schätzung offenbar systematisch zu hoch aus. Dieser Schluss basiert auf dem Vergleich der „tatsächlichen“ Fehlervarianzen $\widehat{Var}_{sim}(\varepsilon)$ [s. die Zahlen in den eckigen Klammern in der Zeile $\widehat{Var}(\varepsilon)$ der Tabelle 2] mit den nach Formel (1.28) geschätzten Fehlervarianzen. Positiv ist hier anzumerken, dass die Schätzungen der Fehlervarianz, und damit der individuellen Standardschätzfehler, bei kleiner Itemzahl zu hoch ausfallen. Damit wird also keine falsche Sicherheit über die Genauigkeit der Schätzer $\hat{\theta}_k$ vorgetäuscht.

Tabelle 2. Ergebnisse der Simulationsstudie

Stichproben-größe	Zahl der Items pro Meßgelegenheit			
	7	15	22	43
	$\widehat{Var}(\xi)$ [1.0]			
1000	.908	.941	.902	.941
2000	.907	.952	.931	.958
5000	.999	.982	.979	1.003
10000	.994	1.000	.996	1.008
	$\widehat{Var}(\hat{\theta})$			
1000	2.309	1.684	1.526	1.396
2000	2.296	1.714	1.539	1.412
5000	2.350	1.753	1.592	1.458
10000	2.334	1.780	1.613	1.465
	$\widehat{Var}(\varepsilon)$			
	[1.075]	[.494]	[.338]	[.171]
1000	1.198	.515	.342	.172
2000	1.197	.517	.343	.172
5000	1.201	.518	.344	.173
10000	1.200	.519	.345	.173
	$\widehat{Var}(\zeta)$ [0.30]			
1000 [.285]	.203	.228	.282	.284
2000 [.291]	.192	.245	.265	.282
5000 [.296]	.150	.254	.269	.282
10000 [.293]	.140	.260	.272	.284

3.4 Zusammenfassung und Diskussion der Simulationsstudie

Zusammenfassend können wir als Ergebnis der Simulationsstudie festhalten, dass die Varianz der Situations-/Interaktionsvariablen bei großer Itemzahl veridikal geschätzt werden kann, dass bei kleiner Itemzahl aber eine deutliche Unterschätzung festzustellen ist, deren Ursache in der Überschätzung der individuellen Standardschätzfehler bei kleinen Itemzahlen liegt. Zu bedenken ist bei der Wertung dieser Ergebnisse allerdings, dass wir kein adaptives Verfahren, sondern die Vorgabe einer festen Itemmenge für alle Personen simuliert haben. Insbesondere ist also beim adaptiven Verfahren zu erwarten, dass die Itemzahl, bei der eine veridikale Schätzung der individuellen Standardschätzfehler zu erwarten ist, wesentlich geringer ist, als bei der hier gewählten fixierten Itemvorgabe.

Die Simulationsstudie betrachtet noch nicht die Besonderheiten beim adaptiven Testen, da allen Personen die gleichen Items vorgegeben werden. Weiterhin ist anzumerken, dass die personenspezifische Varianz aus lediglich 2 oder 3 Messgelegenheiten wahrscheinlich nicht genau genug geschätzt werden kann, sondern nur die über alle Personen gemittelte Varianz. Sie zeigt aber, dass das vorgeschlagene Modell prinzipiell anwendbar ist. Die Güte der Schätzung der Varianz der Situations-/Interaktionsvariablen ζ ist zwar erst bei 43 Items befriedigend. Dabei ist aber zu bedenken, dass wir beim adaptiven Testen schon bei einer erheblich geringeren Itemzahl mit einer hinreichend genauen Schätzung von $Var(\zeta)$ rechnen

können, da dann nur sehr informative Items verwendet werden. Zur Erinnerung: Die Schwierigkeiten der Items in dieser Studie lagen zwischen −3 und + 3 (einschließlich), mit jeweils gleichen Abständen zwischen den Itemschwierigkeiten. Bei einer Person mit einer Fähigkeit von 2 sind daher beispielsweise alle Items mit einer negativen Schwierigkeit – und das ist die Hälfte aller Items – beim hier simulierten Fall mit fester Itemvorgabe fast völlig uninformativ, da sie diese Items mit einer Wahrscheinlichkeit nahe 1 lösen wird.

4. Zusammenfassung und Ausblick

In diesem Bericht haben wir zunächst begründet, warum die Latent-State-Trait-Theorie (LST-Theorie) auch für die wehrpsychologische Eignungsdiagnostik von Bedeutung ist. Das Kernargument dabei ist, dass auch in diesem Bereich der Diagnostik damit gerechnet werden muss, dass die Testergebnisse nicht nur messfehlerbehaftet sind, sondern auch Effekte der psycho-sozio-biologischen *Situation* beinhalten, in der die Probanden getestet werden, ebenso wie Effekte der *Interaktion* zwischen Person und Situation. Die aus einer Erhebung zu einer Messgelegenheit resultierenden Testwerte sind daher in aller Regel unzuverlässiger als die Fehlervarianzen (in der KTT) bzw. die Standardschätzfehler (in der IRT) widerspiegeln. Zur Fehlervarianz muss bei der KTT die Varianz der Situations-/Interaktionsvariablen hinzuaddiert werden und bei der IRT müssen die Standardschätzfehler ebenfalls um den Betrag erhöht werden, der durch die situativ bzw. interaktiv bedingte Variabilität der Fähigkeitszustände zustande kommt.

In einem zweiten Teil wurden dann die Grundbegriffe der LST-Theorie referiert und ein Modell der LST-Theorie dargestellt, das erlaubt, die Varianzen und Varianzanteile der Traitvariablen, Statevariablen, Stateresiduen und Messfehler zu bestimmen. Diese Theorie und die referierten Modelle können dann angewandt werden, wenn zu jeder von mindestens zwei Messgelegenheiten mindestens zwei Messungen (z.B. zwei parallele Tests oder zwei Testhälften) des betrachteten Merkmals vorliegen.

In einem weiteren Teil wurde in diesem Bericht erstmals ein Modell der LST-Theorie konstruiert, das auch für Testwerte, die aus einem CAT-Verfahren resultieren, anwendbar ist. Dabei müssen zwar ebenfalls Messungen zu mindestens zwei Messgelegenheiten vorliegen, aber innerhalb jeder Messgelegenheit genügt ein einziger Testwert (eine einzige Messung), sofern der zugehörige Standardschätzfehler mitgeliefert wird. Das liegt daran, dass mit dem Stichprobenmittelwert der quadrierten Standardschätzfehler die (klassische) Fehlervarianz berechnet werden kann. Eine parallele Messung innerhalb derselben Messgelegenheit zur Ermittlung der Fehlervarianz (oder der Reliabilität) erübrigt sich daher.

In einem letzten Teil wurde mit einer Simulationsstudie untersucht, bei welcher Stichprobengröße und Itemzahl das vorgestellte Verfahren hinreichend genaue Ergebnisse liefert. Dabei zeigte sich, dass die Varianz der Situations-/Interaktionsvariablen bei *großer Itemzahl*

veridikal geschätzt werden kann, dass bei kleiner Itemzahl aber eine deutliche Unterschätzung festzustellen ist, deren Ursache in der Überschätzung der individuellen Standardschätzfehler liegt. Zu Bedenken ist bei der Wertung dieser Ergebnisse allerdings, dass wir kein adaptives Verfahren, sondern die Vorgabe einer festen Itemmenge für alle Personen simuliert haben. Insbesondere ist also beim adaptiven Verfahren zu erwarten, dass die Itemzahl, bei der eine veridikale Schätzung der individuellen Standardschätzfehler zu erwarten ist, wesentlich geringer ist, als bei der hier gewählten fixierten Itemvorgabe

Das vorgestellte Verfahren hat den Vorteil, dass es auch dann anwendbar ist, wenn jede Person im CAT-Verfahren unterschiedliche Items und sogar eine unterschiedliche Zahl von Items zu beantworten hat. Der Nachweis dieser Behauptung durch eine Simulationsstudie, die das CAT-Verfahren simuliert, steht allerdings aus, wird aber als einer der nächsten Schritte in Angriff genommen. Eine solche Simulationsstudie ist insofern aufwendiger, als schon bei der Datengenerierung bei jeder Person mehrfach eine Personenparameterschätzung vorgenommen werden muss, um das informativste nächste Item aussuchen zu können. Weiter bleibt zu untersuchen, inwieweit verfügbare Programme der IRT in der Lage sind, direkt die Varianz der Situations-/Interaktionsvariablen ζ oder gar deren personenbedingte Varianzen zu schätzen. Dies dürfte vor allem deswegen kein leichtes Unterfangen sein, weil im CAT-Verfahren ja jede Person unterschiedliche Items und Itemzahlen zu beantworten hat.

Literatur

Baker, F. B. (1992). *Item Response Theory: Parameter Estimation Techniques.* New York: Marcel Dekker.

Davier, M. von (1997a). *Methoden zur Prüfung probabilistischer Testmodelle.* Kiel: IPN.

Davier, M. von (1997b). WINMIRA – program description and recent enhancements. *Methods of Psychological Research - Online*, *2*, 29-48.

Deinzer, R., Steyer, R., Eid, M., Notz, P., Schwenkmezger, P., Ostendorf, F., & Neubauer, A. (1995). Situational Effects in Psychological Assessment of Traits: The FPI, NEOFFI, and EPI questionnaires. *European Journal of Personality*, *9*, 1-23.

Eid, M. (1995). *Modelle der Messung von Personen in Situationen.* Weinheim: PsychologieVerlagsUnion.

Eid, M. & Hoffmann, L. (1998). Measuring variability and change with an item response model for polytomous variables. *Journal of Educational and Behavioral Statistics*, *23*, 193-215.

Eid, M., Notz, P., Steyer, R. & Schwenkmezger, P. (1993). Validating scales for the assessment of mood level and mood variability by latent state-trait analyses. *Personality and Individual Differences*, *16*, 63-76.

Eid, M., Steyer, R. & Schwenkmezger, P. (1996). Ein Latent-State-Trait-Modell für Variablen mit geordneten Antwortkategorien und seine Anwendung zur Analyse der Variabilitätssensitivität von Stimmungsitems. *Diagnostica*, *42*, 293-312.

Ferring, D., Klauer, T., Filipp, S.-H. & Steyer, R. (1990). Psychometrische Modelle zur Bestimmung von Konsistenz und Spezifität im Bewältigungsverhalten, *Zeitschrift für Differentielle und Diagnostische Psychologie*, *11,* 37-51.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests.* Bern: Huber.

Gähde, U., Jagodzinski, W. & Steyer, R. (1992). On a Structuralist Reconstruction of Latent State-Trait Theory. In H. Westmeyer (Ed.), *The Structuralist Program in Psychology: Foundations and Applications* (pp. 105-119). Toronto: Hogrefe-Huber.

Hambleton, R. K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Boston: Kluwer-Nijhoff.

Hoijtink, H. & Boomsma, A. (1996). Statistical inference based on latent ability estimates. *Psychometrika*, *61*, 1996, 313-330.

Hornke, L. F. (1995). Stand der Technik des Computerunterstützten Adaptiven Testens (CAT). In Bundesministerium der Verteidigung – P II 4 (Hrsg.). *Untersuchungen des Psychologischen Dienstes der Bundeswehr 1993/1995 Band 2* (S. 6-39). München: Verlag für Wehrwissenschaften wVw GmbH.

Kirschbaum, C., Steyer, R., Eid, M., Patalla, U., Hellhammer, D. H. & Schwenkmezger, P. (1990). Cortisol and Behavior: 2. Application of a Latent State-Trait Model to Salivary Cortisol, *Psychoneuroendocrinology*, *15,* 297-307.

Majcen, A.-M., Steyer, R. & Schwenkmezger, P. (1988). Konsistenz und Spezifität bei Eigenschafts- und Zustandsangst. *Zeitschrift für Differentielle und Diagnostische Psychologie, 9,* 105-120.

Reichle, B. & Schmitt, M. (Hrsg.) (1998). *Verantwortung, Gerechtigkeit und Moral.* Weinheim: Juventa.

Rost, J. (1996). *Lehrbuch Testtheorie Testkonstruktion.* Bern: Verlag Hans Huber.

Sands, W. A., Waters, B. K. & McBride, J. R. (Ed., 1997). *Computerized adaptive testing: from enquiry to operation.* Washington, DC: American Psychological Association.

Schmitt, M. (1990). *Konsistenz als Persönlichkeitseigenschaft?* Berlin: Springer.

Schmitt, M. J., Schwartz, S. H., Steyer, R. & Schmitt, T. (1993). Measurement models for the Schwartz Values Inventory. *European Journal of Psychological Assessment*, *9*, 107-121.

Schmitt, M. & Steyer, R. (1993). A Latent State-Trait Model (not only) for Social Desirability, *Personality and Individual Differences*, *14*, 519-529.

Steyer, R. (1995). Was ist eine Persönlichkeitseigenschaft? In K. Pawlik (Hrsg.), *Bericht über den 39. Kongreß der Deutschen Gesellschaft für Psychologie in Hamburg 1994* (S. 720-725). Göttingen: Hogrefe.

Steyer, R. & Eid, M. (1993). *Messen und Testen*. Berlin: Springer.

Steyer, R., Eid, M. & Schwenkmezger, P. (1997). Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research Online*, *2*, 21-33.

Steyer, R., Ferring, D., & Schmitt, M. J. (1992). States and Traits in Psychological Assessment. *European Journal of Psychological Assessment*, *8*, 79-98.

Steyer, R., Partchev, I. & Shanahan, M. J. (2000). Modeling true intra-individual change in structural equations models: The case of poverty and children´s psychosocial adjustment. In: T. D. Little & K. U. Schnabel (eds.), *Modeling longitudinal and multilevel data* (pp. 109-126). Mahwah, NJ: Lawrence Erlbaum.

Steyer, R. & Schmitt, M. J. (1990a). Latent State-Trait Models in Attitude Research. *Quality and Quantity, 24*, 427-445.

Steyer, R. & Schmitt, M. (1990b). The Effects of Aggregation Across and Within Occasions on Consistency, Specificity, and Reliability. *Methodika, 4,* 58-94.

Steyer, R., Schwenkmezger, P. & Auer, A. (1990). The Emotional and Cognitive Components of Trait Anxiety: A Latent State-Trait Anxiety Model. *Personality and Individual Differences, 11,* 125-134.

Steyer, R., Schmitt, M. & Eid, M. (1999). Latent State-Trait Theory and Research in Personality and Individual Differences. *European Journal of Personality*, *13*, 389-408.

Storm, E. G. (1999). *Adaptive (leistungsabhängige) Testverfahren – Hintergründe, Testökonomie und Akzeptanz.* Unveröffentlichter Arbeitsbericht aus dem Psychologischen Dienst der Bundeswehr. Bonn.

Wainer, H. et al. (1990) (Eds.). *Computerized adaptive testing: a primer.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Anhang: Quelltext und Beschreibung des SYSTAT-Basic Programms zur Datenerzeugung in der Simulationsstudie

basic

Dieser Befehl startet das Systat BASIC Programm.

new

Mit diesem Befehl wird ein neuer Datensatz erzeugt.

repeat 10000

Dieser Befehl reserviert 10000 Zeilen für die Zufallszahlen.

let ksi=zrn

Dieser Befehl definiert ksi als standardnormalverteilte [N(0, 1)] latente Trait-Variable.

let sigmazeta=sqr(0.3*xrn(1))

Mit diesem Befehl wird sigmazeta (die personspezifische Standardabweichung über Messgelegenheiten) als Quadratwurzel einer mit 0.3 multiplizierten χ^2-verteilten Zufallsgröße (mit einem Freiheitsgrad) definiert.

dim theta(3)

Dieser Befehl reserviert drei Variablen für die drei Messgelegenheiten.

for i=1 to 3
let theta(i)=ksi+zrn(0,sigmazeta)

Mit theta (1 bis 3) werden die drei latenten State-Variablen bezeichnet. Jede einzelne dieser drei Variablen setzt sich aus dem ksi-Wert und dem Wert einer normalverteilten Zufallsgröße Zeta mit dem Mittelwert null und der personspezifischen Standardabweichung sigmazeta zusammen. Die Werte von Zeta repräsentieren die Situations/ Interaktionseffekte

next

dim it(129)

Dieser Befehl reserviert drei mal 43 Variablen für die drei Antwortmuster.

```
for j=1 to 3
  for i=1 to 43
    let it((j-1)*43+i) = 1/(1+exp(((i-22)/7)-theta(j))) > urn
  next
```

Mit den ersten beiden Befehlszeilen werden Schleifen über alle drei Messgelegenheiten und alle 43 Items definiert. Der Ausdruck (j-1)*43+i zeigt auf ein Item (1 ...129). Der Ausdruck (i-22)/7 ergibt die Itemschwierigkeit. (Für das erste Item i=1 ergibt sich eine Schwierigkeit von –3, für das letzte Item i=129 eine Schwierigkeit von +3, so dass die Schwierigkeiten insgesamt im Bereich von –3 bis +3 liegen.) Der Ausdruck 1/(1+exp((i-22)/7)-theta(j))) ist die Wahrscheinlichkeit für eine richtige Antwort gegeben die Schwierigkeit (i-22)/7 und die Fähigkeit theta(j). Der Ausdruck urn ist eine gleichverteilte (0,1) Zufallsgröße. Der Gesamtausdruck 1/(1+exp(((i-22)/7)-theta(j))) > urn ergibt somit 0, wenn falsch und 1, wenn richtig geantwortet wurde.

next
run

Die mit diesem Programm erzeugten Daten finden sich auf unserer Homepage unter http://www.uni-jena.de/svw/metheval/daten/start.html