Technische Universität Berlin

Fakultät I
Institut für Sprache und Kommunikation
Fachgebiet Audiokommunikation

Vergleichende Simulation adaptiver, psychometrischer Verfahren zur Schätzung von Wahrnehmungsschwellen

Magisterarbeit

vorgelegt von

Stefanie Otto

Matrikelnummer 221462
geboren am 1. September 1983 in Strausberg

Erstgutachter: Prof. Dr. Stefan Weinzierl
Zweitgutachter: Dr. Hans-Joachim Maempel

7. August 2008

Die selbstständige Anfertigung dieser Arbeit versichere ich an Eides statt.

Berlin, den 7. August 2008, _____

Stefanie Otto

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Fragestellung

Eine häufige Problemstellung psychophysischer Untersuchungen ist die experimentelle Bestimmung der Reizschwelle (*threshold*) einer bestimmten menschlichen Wahrnehmung. Die absolute Reizschwelle dient als gemeinsamer Bezugspunkt von Reiz und Empfindung und bildet zusammen mit der Unterschiedsschwelle (*just noticeable difference*, JND) die Grundlage von Empfindungsskalen. Anwendung finden die Ergebnisse nicht nur in der allgemeinen Wahrnehmungsforschung und im klinischen Bereich, sondern auch als Maßstab für die Entwicklung von Informationsübertragungssystemen so z.B. bei der Erkennung und Bewertung verschiedener Qualitäten der Audio-Codierung.

Gustav Theodor Fechner entwickelte 1860 eine Theorie zur Messung innerer Empfindungen und erarbeitete verschiedene Methoden wie das Herstellungsverfahren, sowie das Grenzverfahren und das Konstanzverfahren zur Bestimmung von absoluten und Unterschiedsschwellen [4]. Diese experimentellen Verfahren gründen auf der Idee, dass die Versuchsperson beim Zutreffen einer inneren Bedingung eine bestimmte Operation ausführen soll (z.B. Knopf drücken, Ja/Nein sagen, Regler verstellen). Dazu wird eine mutmaßliche Einflussgröße (unabhängige Variable) gezielt durch Manipulation oder Selektion verändert und die Ausprägung einer anderen Größe (abhängige Variable) gemessen. Trefferquoten über der Bedingungsvariation aufgetragen, liefern eine Schätzung der psychometrischen Funktion, aus der weitere Parameter berechnet werden können (z.B. der Punkt subjektiver Gleichheit zweier Reize bei 50% Trefferquote). Bei der Auswertung der gewonnenen Daten muss jedoch der Einfluss anderer Größen (moderierende Variablen) berücksichtigt werden [16].

Die klassischen Verfahren weisen einige Nachteile auf. Sie sind oft sehr zeitaufwendig, da die angebotenen Stimuli anfangs weit entfernt von der Schwelle liegen. Die Bewertung dieser Reize enthält somit wenig Information über die Lage der wahren Schwelle [6]. Ein weiteres Problem ist, dass sensorische Bewertungen durch Versuchspersonen unausweichlich mit einem individuell unterschiedlichen Entscheidungskriterium konfundiert sind [16]. Um dieses Kriterienproblem zu lösen und die bestehenden Messverfahren bezüglich Genauigkeit und Effizienz weiterhin zu verbessern, wurden in den letzten 60 Jahren zahlreiche Modifikationen und Erweiterungen entwickelt.

Moderne, adaptive Verfahren reagieren auf die Erkenntnisleistung der Versuchsperson, indem sie die Stärke des Reizes entsprechend einer adaptiven Regel anpassen. Dadurch arbeiten sie effizienter und auch genauer als klassische Verfahren mit vorbestimmtem Versuchsablauf. Außerdem versucht man durch verschiedene Abfragedesigns objektiv richtige und falsche Antworten zu erhalten und somit die Störeinflüsse des Entscheidungskriteriums zu minimieren [6].

Die existierenden Varianten adaptiver, psychometrischer Verfahren sind vielfältig. Sie unterscheiden sich nicht nur in ihrer Herangehensweise zu Ablauf und Abfragedesign, sondern auch in den Vorannahmen über die zugrunde liegende, psychometrische Funktion [26]. Deshalb sind allgemeingültige Aussagen zu ihrer Eignung nicht möglich. In der vorliegenden Arbeit werden einige der wichtigsten, adaptiven Verfahren implementiert und mittels Simulationen auf ihre Eigenschaften hin untersucht. Weiterhin werden Stabilität, Genauigkeit und Effizienz der Prozeduren analysiert und mit bestehenden Ergebnissen aus der Literatur verglichen. Ziel ist es einen Überblick über adaptive, psychometrische Verfahren zu bieten, Vor- und Nachteile herauszustellen und Empfehlungen für den Einsatz in der Praxis zu geben.

1.2 Grundlagen

Die seit 1960 entwickelten, modernen Verfahren waren anfangs meist Modifikationen der klassischen Konstanz- und Grenzverfahren. Mit der Entwicklung der Theorien zum Entscheidungsverhalten (Luce, 1959 [13] und 1963 [14]) und zur

Signalentdeckung (Green & Swets, 1966 [7]) wurde eine Basis für neue psychometrische Verfahren geschaffen. Seit den 80er Jahren wird die Durchführung und Auswertung der Tests durch den Einsatz von Computern zunehmend erleichtert. Zum Beispiel ist die adaptive Anpassung durch online generierte Stimuli erst in computergesteuerten Evaluationen möglich geworden.

1.2.1 Adaptive Reizdarbietung

Bei adaptiven Verfahren sind die Reizstärken (und Schrittweiten) nicht schon vor Beginn der Messung fixiert, sondern werden einer adaptiven Regel folgend und abhängig von der Antwort der Versuchsperson variiert. Die Darbietung der Reize abhängig von vorangehenden Antworten wird als stationärer stochastischer Prozess betrachtet. Dies impliziert, dass aufeinander folgende Antworten als statistisch unabhängig betrachtet werden. Da die Antworten einen stochastischen Prozess darstellen (Zufallsvariable Z_n), ist die Reizdarbietung demzufolge auch ein stochastischer Prozess (Zufallsvariable X_n). Mögliche Werte sind $z_i = 1$ für positive (korrekte) und $z_i = 0$ für negative Antworten. Unter der Annahme, dass die Antworten der Versuchsperson bei jedem festen Stimulus Level binomial verteilt sind, gilt dann für die psychometrische Funktion $\psi(x)$:

$$\begin{aligned} Prob\{Z_n = 1|X_n\} &= \psi(X_n) \\ Prob\{Z_n = 0|X_n\} &= 1 - \psi(X_n) \end{aligned} \tag{1.1}$$

Jedes adaptive Verfahren $\mathcal{A}$ kombiniert also die Reizdarbietungen X_n und die korrespondierenden Antworten Z_n des n-ten Trials und der vorangehenden Trials mit dem Konvergenzniveau ϕ um eine optimale Reizintensität X_{n+1} zu ermitteln, die als nächstes dargeboten wird:

$$X_{n+1} = \mathcal{A}\{\phi, n, X_n, Z_n, \ldots, X_1, Z_1\} \tag{1.2}$$

Diese Vorgehensweise konzentriert die Messungen auf den Bereich der größten Unsicherheit und konvergiert im Idealfall sehr schnell gegen einen bestimmten Prozent-Korrekt-Wert. Deshalb sind adaptive Verfahren generell effizienter als klassische

Methoden der Schwellenbestimmung. Gleichzeitig minimiert eine kürzere Messzeit auch die Einflüsse, die durch Veränderungen des Entscheidungskriteriums der Versuchsperson zustande kommen können [19].

1.2.2 Paradigmen

In klassischen Ja-Nein-Abfragen muss die Versuchsperson entscheiden, ob sie zwei unterschiedliche Reize auch unterschiedlich wahrnimmt (Ja-Antwort) oder keinen Unterschied erkennt (Nein-Antwort). Im Gegensatz dazu stehen die sogenannten Forced-Choice-Paradigmen, die auf die Signalentdeckungstheorie zurückgehen. Dabei muss die Versuchsperson einen Zielreiz aus mehreren zeitlich oder örtlich verteilten Reizen auswählen. Kann die Versuchsperson keinen Unterschied feststellen, ist sie in diesem Fall gezwungen zu raten. Der Vorteil dieser erzwungenen Zuweisungen ist, dass man objektiv richtige und objektiv falsche Antworten erhält. Sie ermöglichen also eine Betrachtung des sensorischen Anteils einer Antwort getrennt von dem individuell verschiedenen Entscheidungskriterium. Es bestehen keine Vorgaben dazu, welche psychometrischen Methoden mit welchem Paradigma kombiniert werden dürfen. Allerdings sind einige Verfahren besser geeignet, die höhere Ratewahrscheinlichkeit zu tolerieren.

1.2.3 Psychometrische Funktion und Threshold Definition

Die psychometrische Funktion $\psi(x)$ ist eine Verteilungsfunktion der kumulierten Wahrscheinlichkeiten binärer Antworten. Sie stellt die Summe der positiven Antworten (Ja bzw. korrekt) aufgetragen über der Reizintensität dar. Ergebnisse mehrerer Messungen mit gleichem Stimulus-Setup aufgetragen über der Reizintensität ergeben die psychometrische Funktion. Im Idealfall ist dies eine Kurve, die mit wachsender Reizintensität von der Wahrscheinlichkeit 0 auf 1 monoton ansteigt. Das Problem bei realen Versuchspersonen ist jedoch, dass sogar bei sehr hohen Reizintensitäten der Stimulus nicht immer richtig erkannt wird. Dieses Phänomen bezeichnet man als Lapsing und der Prozentsatz der beobachteten Verpasser (auch false negative errors) ergibt die Lapsing Rate, die meist bei $p_l = 2 - 3$ % liegt.

Außerdem kommt es vor, dass die Versuchsperson bei unterschwelligen Reizen durch Raten richtig liegt. In Ja-Nein-Verfahren wird dies als Rauschen interpretiert, bei Forced Choice-Verfahren ist die Versuchsperson jedoch gezwungen zu raten. Bei n Alternativen ergibt sich die Ratewahrscheinlichkeit zu $p_g = 1/n$, wobei Versuchspersonen vorausgesetzt werden, die beim Raten keine der Alternativen bevorzugen. Treten p_g und p_l als Asymptoten der psychometrischen Funktion auf, wird eine Korrektur des Konvergenzniveaus nötig. Das heißt, dass die theoretischen Prozent-Korrekt-Werte an die tatsächlich gemessenen Werte $\psi(x)$ angepasst werden müssen. Man bedient sich dazu der Formel von Abbott [26]:

$$\psi^*(x) = \frac{\psi(x) - p_g}{1 - p_g - p_l} \quad \text{oder auch} \quad \psi(x) = p_g + (1 - p_g - p_l)\psi^*(x) \tag{1.3}$$

Die Varianz der Prozent-Korrekt-Werte (und damit auch die Genauigkeit der Messung) ist abhängig von Anzahl n der Trials bei einem bestimmten Stimulus Level und von der unbekannten „wahren“ Antwortwahrscheinlichkeit. Die Varianz der Wahrscheinlichkeit einer positiven Antwort p_s bei binomial verteilten Antworten ist dann:

$$Var(p_s) = p_s(1 - p_s)/n \tag{1.4}$$

Sollen die erhobenen Daten nach der Messung an ein theoretisches Modell (also eine bestimmte psychometrische Funktion) angepasst werden, ist es deshalb empfehlenswert die Varianz der Rate der unkorrigierten Wahrscheinlichkeiten als Gewichtungsfaktor für die korrigierten Wahrscheinlichkeiten zu nutzen. Für $p = 0.5$ wird die Varianz maximal und die benötigte Anzahl von Messungen ist an diesem Punkt minimal. Soll also ein anderes Konvergenzniveau erreicht werden als der 50%-Korrekt-Wert, muss die Anzahl der Trials dementsprechend erhöht werden.
In einigen adaptiven Verfahren wird eine bestimmte Verteilungsfunktion als psychometrisches Modell der zu untersuchenden Wahrnehmung zugrunde gelegt. Die somit festgelegte Form der psychometrische Funktion reduziert die mögliche Anzahl an Schwellwerten und erleichtert somit die Berechnungen während der Messung. Oft verwendete Formen sind die der Gauss-Verteilung, die logistische und die Weibull-Funktion.

Als Schwelle (*threshold* θ) einer bestimmten Wahrnehmung wird die Reizstärke x_ϕ bezeichnet, die bei einer bestimmten Anzahl der Darbietungen (z.B. 75%) korrekt erkannt wird. Der Prozent-Korrekt-Wert ϕ kann theoretisch frei gewählt werden. Bei Ja-Nein-Abfragen wird oft der Stimulus als Schwelle definiert, bei dem positive und negative Antworten gleichhäufig sind (50%-Korrekt-Wert, auch Punkt subjektiver Gleichheit). Bei Forced-Choice-Paradigmen wird meist die Mitte des Intervalls zwischen Lapsing Rate und Guessing Rate gewählt: $\phi = (1 - p_l + p_g)/2$

1.2.4 Evaluation psychometrischer Verfahren

Psychometrische Tests werden je nach Anforderung der jeweiligen Untersuchung konzipiert. Das Verfahren soll dabei möglichst den drei Testgütekriterien Objektivität, Validität und Reliabilität genügen. Gleichzeitig soll so effizient wie möglich getestet werden, um auch bei wenigen Versuchspersonen viele, aussagekräftige Schwellwerte ermitteln zu können. Psychometrische Verfahren sollten demnach hinsichtlich Aufwand und Nutzen bewertet werden. Die empirisch ermittelte Schwelle ist ein statistischer Wert, dessen Gültigkeit durch verschiedene Größen bestimmt wird:

- systematische Fehler (*bias*)
- Genauigkeit der Messung (zufällige Fehler, Varianz)
- Effizienz (Verhältnis von Anzahl der Messungen zur Genauigkeit)

Die Bestimmung des systematischen Fehlers stellt bei realen Messungen ein kaum zu lösendes Problem dar. Gesucht ist dabei die Differenz des Schätzwertes zum wahren Wert der Schwelle. Der Bias einer Prozedur kann deshalb nur in Simulationen bestimmt werden, in denen der wahre Schwellwert bekannt ist.
Der Bias bei r gemessenen Schwellwerten $\hat{\theta}_r$ wird dann wie folgt definiert:

$$b_{\hat{\theta}} = \frac{1}{r} \sum_r (\theta_{true} - \hat{\theta}_r) = \theta_{true} - \mu_{\hat{\theta}} \tag{1.5}$$

Die Genauigkeit κ_θ eines Verfahrens kann als Inverse der Varianz des ermittelten Schwellwerts $\hat{\theta}$ bestimmt werden:

$$\kappa_\theta = \frac{1}{\sigma^2_{\hat{\theta}}} = \frac{r-1}{\sum\limits_r (\hat{\theta}_r - \mu_{\hat{\theta}})^2} \tag{1.6}$$

Zur Messung der Effizienz haben Taylor und Creelman (1967) und Taylor (1971) den so genannten *sweat factor* K als Produkt der Varianz der ermittelten Schwellwerte $\sigma^2_{\hat{\theta}}$ und der Anzahl der Trials n definiert [25]:

$$K = n\sigma^2_{\hat{\theta}} = n\frac{\sum\limits_r (\mu_{\hat{\theta}} - \hat{\theta}_r)^2}{r-1} \tag{1.7}$$

Der *sweat factor* repräsentiert den Aufwand, der bei einer bestimmten Messung betrieben wird und dient somit der vergleichenden Beurteilung verschiedener psychometrischer Verfahren. Es zeigt sich, dass die Anzahl der Trials n und die Genauigkeit κ_θ invers miteinander verknüpft sind. Deshalb muss bei der Wahl des Verfahrens ein geeigneter Kompromiss zwischen Aufwand und Nutzen gefunden werden [27]. Um absolute Werte der Effizienz zu erhalten, müsste eine ideale Prozedur als Referenz existieren. Taylor (1971) schlägt dafür die Messung der asymptotischen Varianz des Robbins-Monro-Verfahrens vor [24]. Eine wichtige Einflussgröße der Effizienz ist das Verhalten der Prozedur bei verschiedenen Reizintensitäten zu Beginn der Messung. Daraus folgt auch, dass der anfängliche Schätzwert der Schwelle einen bedeutenden Einfluss auf die Varianz der ermittelten Schwellwerte haben kann.

1.2.5 Kategorisierung von Verfahren

Um eine bessere Übersicht über die verschiedenen Varianten psychometrischer Verfahren zu gewinnen, wird die folgende Strukturierung nach Marvit et al. (2003) vorgeschlagen (siehe Abb. 1.1). Danach werden psychometrische Verfahren im Allgemeinen durch zwei Aspekte charakterisiert, die theoretisch unabhängig, jedoch in der Praxis eng miteinander verwoben sind.

Ein Aspekt ist das *Paradigma*, welches einerseits die *Art der Reizdarbietung* und andererseits die *Aufgabe der Versuchsperson* beschreibt. Bei der Art der Reizdarbietung unterscheidet man weiterhin die Anzahl der Intervalle in einem Trial und den Inhalt dieser Intervalle. Den zweiten Aspekt psychometrischer Verfahren bildet die *Methode* (angelehnt z.B. an Fechners Methode des konstanten Reizes). Die Methode im engeren Sinn beinhaltet die *Strategie* zur Platzierung der Stimuli und die Berechnungsvorschrift (*datum definition*). Die Strategie wiederum manifestiert sich in drei Regeln: 1) *starting rule*, die den Reiz für den ersten Trial bestimmt, 2) *progression rule*, welche die Wahl des jeweils nächsten Reizes vorgibt, und 3) *stopping rule*, die festlegt, wann eine Messreihe beendet wird. Die Vorschrift zur Berechnung gibt an, wie der resultierende Parameter aus der gegebenen Reihenfolge der Reizdarbietungen und den entsprechenden Antworten der Versuchspersonen berechnet wird [17].

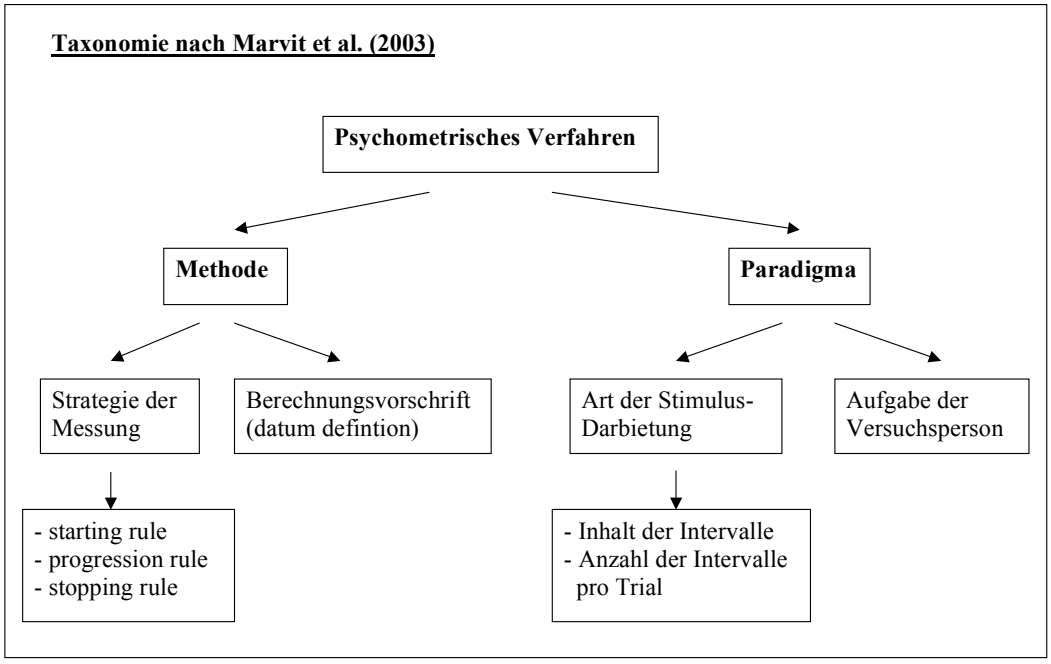

Abbildung 1.1: Strukturierung psychometrischer Verfahren

Die verschiedenen Phasen eines adaptiven Verfahrens, Stimulusplatzierung, Abbruchkriterium und Berechnung des Schwellwerts, sind generell unabhängig voneinander wählbar. Dadurch ergeben sich viele Kombinationsmöglichkeiten und damit die große

Vielfalt an psychometrischen Verfahren. Ein Überblick über die existierenden Varianten findet sich bei Treutwein [26]. Dieser unterscheidet hauptsächlich zwei Kategorien, parametrische und nichtparametrische Verfahren, je nachdem, ob und inwiefern Annahmen über die zugrunde liegende psychometrische Funktion in die Messung miteinbezogen werden.

Im Kontext dieser Arbeit werden nur Verfahren untersucht, bei denen die Antworten der Versuchspersonen auf binäre Äußerungen (Ja/Nein, Korrekt/Falsch) reduziert werden können und bei denen die Variation der Stimuli ein eindimensionales Kontinuum repräsentiert. Außerdem werden nur die Verfahren einbezogen, die es sich zum Ziel setzen die Wahrnehmungsschwelle zu bestimmen. Bei der Lösung von mehrdimensionalen Fragestellungen (z.B. Bestimmung der Lage und Steigung der psychometrischen Funktion) treten weitere Probleme auf, die den Rahmen dieser Arbeit überschreiten würden.

1.2.6 Staircase-Verfahren

Die *Staircase*-Prozeduren stellen nichtparametrische Verfahren dar. Das bedeutet, dass die Form der psychometrischen Funktion vor der Messung nicht bekannt sein muss. Einzige Annahme ist, dass die Funktion mit wachsender Reizstärke monoton steigt. Somit wäre eine Schätzung von Form, Steigung und Schwellwert möglich. Die Messung beginnt bei einem *educated guess*, wobei die Anfangsstimuli oft in einem bestimmten Bereich um die vermutete Schwelle randomisiert werden. Die Anpassung der Reizintensität erfolgt mit fester Schrittgröße δ und nach jedem Trial. Ein Wechsel der Schrittrichtung wird vorgenommen, wenn sich die Antwortkategorie verändert.

$$X_{n+1} = X_n - \delta(2Z_n - 1) \tag{1.8}$$

Die Messreihe endet nach einer bestimmten Anzahl von Trials oder Umkehrungen (*turnarounds*). Der Schätzwert der Schwelle wird durch Mittelung der Reizstärken an den Umkehrpunkten ermittelt, wobei der erste *turnaround* oft vernachlässigt wird [2].

Ein Problem von Staircase-Verfahren ist, neben der Wahl des zu untersuchenden Bereichs der Reizstärken (*Range*), die Wahl geeigneter Schrittgrößen (also der Auflösung). Mit großen Schrittweiten wird der Bereich der Schwelle schnell erreicht, jedoch sind die ermittelten Schätzwerte nicht sehr genau. Kleine Schrittgrößen führen zu einer höheren Genauigkeit, benötigen aber weitaus mehr Trials um zu konvergieren. Nach Dixon und Mood (1948) soll die Schrittweite abhängig von der Varianz im Bereich von 0.5σ bis 2.4σ gewählt werden [2]. Die Varianz ist jedoch in den wenigsten Fällen schon vor der Messung bekannt. Andere Autoren schlagen vor die Schrittgröße nach einer bestimmten Anzahl von Umkehrungen zu halbieren [11].
Nachteil der Standard Staircase-Methode ist, dass sie stets bei 50% korrekter Zuweisung konvergiert. Eine Möglichkeit der Anpassung des Konvergenzniveaus bieten die *Transformed Up-Down*-Verfahren nach Levitt (1971) [11]. Als Beispiel wird hier die *2Down-1Up*-Regel erläutert. Dabei wird die Reizstärke nach einer falschen Zuweisung $(-)$ erhöht, jedoch erst nach zwei richtigen Antworten $(++)$ gesenkt. Nach nur einer richtigen Zuweisung $(+)$ wird die Reizstärke nicht verändert. Die Wahrscheinlichkeiten für die *Down*- bzw. *Up*-Sequenz ergeben sich zu:

$$p(++) = p(x)^2 \quad \text{und} \quad p(- \text{ oder } + -) = p(x)[1 - p(x)] + [1 - p(x)] \tag{1.9}$$

Diese Prozedur konvergiert auf dem Niveau, bei dem Abwärtsschritt $(++)$ und Aufwärtsschritt $(-)$ oder $(+-)$ gleichwahrscheinlich sind.

$$p_{Konvergenz} = \sqrt{1/2} = 0.707$$

Von Levitt wurde eine Reihe von adaptiven Regeln vorgeschlagen, um verschiedene Konvergenzniveaus nach der Transformed Up-Down-Methode zu erreichen. Eine andere Alternative stellt die Methode der Stochastischen Approximation (Robbins & Monro, 1951) dar, wobei ein beliebiges Konvergenzniveau erzielt werden kann [21].

1.2.7 PEST und Modifikationen

PEST steht für *Parameter Estimation by Sequential Testing* und wurde von Taylor und Creelman (1967) entwickelt. Es ist eine der ersten Methoden, die Erkenntnisse der sequentiellen Statistik in die Messung mit einbeziehen. Dabei wird ein vereinfachter *sequential probability ratio test* (SPRT, nach Wald, 1947) angewandt, um zu ermitteln, ob die aktuelle Reizintensität verändert werden muss. Dazu wird die Anzahl der korrekten Antworten $N(c)$ und die Anzahl der Trails n_x bei der aktuellen Reizstärke x gezählt. Die erwartete Anzahl korrekter Antworten nach n_x Stimuluspräsentationen und bei gewünschtem Konvergenzniveau ϕ ergibt sich aus dem Mittelwert der Binomialverteilung $\mathcal{B}(n_x, \phi)$:

$$E[N(c)] = \phi n_x \tag{1.10}$$

Der Versuchsleiter definiert mittels der Konstante W (*deviation limit*) eine obere und untere Grenze für $N(c)$. Taylor & Creelman schlagen einen Wert von $W = 1$ für $\phi = 0.75$ vor [25].

$$N_b(c) = E[N(c)] \pm W \tag{1.11}$$

Liegt die tatsächlich beobachtete Anzahl korrekter Antworten $N(c)$ innerhalb dieser Grenzen, wird die Messung bei dieser Reizstärke fortgeführt. Liegt $N(c)$ jedoch außerhalb des Intervalls, kann die Nullhypothese, die besagt, dass der Teststimulus dem Schwellwert entspricht, widerlegt werden. Dann wird die aktuelle Reizstärke dementsprechend erhöht oder gesenkt. Die Schrittgröße wird anhand eines Sets heuristischer Regeln bestimmt:

1. Halbierung der Schrittgröße bei jedem *Reversal* (Umkehrung der Schrittrichtung)
2. Der zweite Schritt in gleicher Richtung erfolgt mit gleicher Schrittgröße wie der erste.
3. Der vierte und jeder weitere Schritt in gleicher Richtung ist doppelt so groß wie der vorangehende.
4. Der dritte aufeinander folgende Schritt in gleicher Richtung wird nur dann in der Größe verdoppelt, wenn bei dem Schritt vor dem letzten Reversal keine Verdopplung stattfand.

Zu Beginn der Messung müssen also nicht nur der Range, das Konvergenzniveau ϕ und die anfängliche Schrittgröße gewählt werden, sondern auch das Deviation Limit W. Die Messung endet, wenn eine vorgegebene minimale Schrittgröße erreicht wird. Zur Berechnung des resultierenden Schwellwerts existieren verschiedene Ansätze. Sie erfolgt zum Beispiel durch Mittelung (*RAT Mode*) oder man wählt den zuletzt ermittelten Stimulus als besten Schätzwert (*MOUSE Mode*) [26]. Findlay lieferte 1978 mit dem *More Virulent PEST*-Verfahren eine weitere Modifikation, die die Messung beschleunigen soll. Dabei ist das Deviation Limit W nicht fixiert, sondern variiert abhängig von der Anzahl der Trials und Reversals [5].

1.2.8 Maximum Likelihood und Bayes-Schätzung

Maximum Likelihood (kurz ML) und *Bayes*-Prozeduren sind parametrische Verfahren, die schon vor der Messung die Form und meist auch die Steigung einer bestimmten psychometrischen Funktion als bekannt voraussetzen. Dies erleichtert die Bestimmung der Schwelle insofern, als dass sich mögliche Varianten der psychometrischen Funktion auf eine Kurvenschar reduzieren, die nur noch entlang der Stimulus-Achse verschoben wird. Außerdem wird die Genauigkeit der Prozedur verbessert, wenn die Guessing Rate p_g und die Lapsing Rate p_l in die zugrunde liegende psychometrische Funktion miteinbezogen werden.

Grundlage dieser Verfahren bilden die Theorie der statistischen Schätzung und die statistische Entscheidungstheorie, die sich damit beschäftigen aus einem gegebenen Set von Daten den gewünschten Populationsparameter zu ermitteln. In der Psychophysik werden dazu die Maximum-Likelihood- und die Bayes-Schätzung angewandt, um mit einer möglichst kleinen Anzahl von Trials und optimal platzierten Stimuli einen validen Schätzwert zu gewinnen. Dabei wird aus der Likelihood-Funktion oder der *posterior probability density function* (*posterior pdf*) der beste aktuelle Schätzwert für die Schwelle berechnet und als nächster Stimulus dargeboten. Der Schwellwert liegt meist im Wendepunkt der psychometrischen Funktion, wo die Steigung und damit auch die Varianz der kumulierten Wahrscheinlichkeitsverteilung maximal ist.

Richtige wie auch falsche Zuweisungen von Stimuli im Bereich des Wendepunkts haben demnach höchsten Informationsgehalt. Die Konzentration der Messung bei der vermuteten Schwelle soll also den Informationsgewinn weiter maximieren.

Prinzip der Maximum Likelihood-Schätzung

Hinter dem Prinzip der Maximum Likelihood-Schätzung steht die Idee, dass verschiedene Populationen unterschiedliche Ergebnisse erzeugen. Jedes einzelne Messergebnisse stammt mit höherer Wahrscheinlichkeit von der einen als von der anderen Population. Zur Schätzung des Populationsparameters θ wird aus der Schar möglicher psychometrischer Funktionen diejenige gewählt, die beim gegebenen Stimulusverlauf mit größter Wahrscheinlichkeit zu den erhaltenen Antworten führt. Diese Wahrscheinlichkeit kann mittels der *conditional joint probability density function* $f(\mathbf{x}|\theta)$ berechnet werden. Sie beschreibt die Wahrscheinlichkeit, dass bei einem gegebenen Set von bisher dargebotenen Stimuli X_n und korrespondierenden Antworten Z_n genau dieser Wert x die Schwelle darstellt. Der Maximum Likelihood-Schätzwert ist der Reiz, für den die conditional joint pdf $f(\mathbf{x}|\theta)$ ihr Maximum erreicht. Deshalb wird diese Dichtefunktion nicht normalisierter Wahrscheinlichkeiten auch *Likelihood-Funktion* $\mathcal{L}(\theta) = f(\mathbf{x}|\theta)$ genannt. Zur Vereinfachung wird angenommen, dass die psychometrische Funktion invariant bezogen auf die x-Achse ist. Der Vektor möglicher Schwellwerte θ_i ist somit identisch mit dem Vektor möglicher Stimuli x_i. Nach n Stimuluspräsentationen bei den Reizintensitäten $\mathbf{x}$ ergibt sich die Likelihood Funktion als Produkt der Einzelwahrscheinlichkeiten über dem Vektor möglicher Schwellwerte.

$$\begin{aligned} \mathcal{L}(\theta|x_1 \dots x_n) &= p(\theta|x) f(\mathbf{x}) = \prod_{i=1}^{n} \mathcal{L}(\theta|x_i) \\ &= \mathcal{L}(\theta|x_n) \prod_{i=1}^{n-1} \mathcal{L}(\theta|x_i) \end{aligned} \tag{1.12}$$

Die Wahrscheinlichkeit einer bestimmten Antwort wird mithilfe der angenommenen psychometrischen Funktion berechnet. Der Wert der Likelihood-Funktion $\mathcal{L}(\theta|x_i)$ für Trial i ergibt dann:

$$\mathcal{L}(\theta|x_i) = \begin{cases} p_+(x_i,\theta) = \psi(x_i,\theta) \\ p_-(x_i,\theta) = 1 - \psi(x_i,\theta) \end{cases} \tag{1.13}$$

Prinzip der Bayes-Schätzung

Die Bayes-Schätzung nutzt die Vorteile von *a priori* Informationen. Schon vorhandene Informationen (z.B. aus Vorversuchen) über die Wahrscheinlichkeitsverteilung der möglichen Schwellwerte θ werden durch die *prior probability density function (prior pdf)* $g(\theta)$ ausgedrückt. Diese Strategie führt zur besseren Platzierung der Stimuli und vermeidet zu große Schritte in nicht interessierende Bereiche während der ersten Trials. Bei einer Gauss-förmigen prior pdf wird der erste zu testende Stimulus demnach in der Mitte des Range liegen. Ist kein Vorwissen vorhanden, wird eine nichtinformative Verteilung genutzt (z.B. Rechteckverteilung, jeder Wert θ ist gleichwahrscheinlich).
Im Unterschied zur ML-Schätzung sucht man bei der Bayes-Methode nicht nach einem festen Parameter, sondern nach einer Wahrscheinlichkeitsverteilung für die möglichen Schwellwerte. Aus der prior pdf $g(\theta)$ und den gewonnenen Daten aus vorangehenden Trials $f(\mathbf{x}|\theta)$ ergibt sich die posterior pdf nach dem *Bayes-Theorem* (ausführlicher siehe bei Watson & Pelli [27]).

$$p_{post}(\theta|x) = \frac{f(\mathbf{x}|\theta)g(\theta)}{h(\mathbf{x})} \tag{1.14}$$

$h(\mathbf{x})$ stellt einen konstanten Normalisierungsfaktor dar, der nur von vom Set der möglichen Stimuli $\mathbf{x}$ abhängig ist. Die nicht normalisierte posterior pdf kann auch umgeschrieben werden zu:

$$\begin{aligned} \mathcal{L}(\theta) &= p_{post}(\theta|x)h(\mathbf{X}) \\ &= f(\mathbf{x}|\theta)g(\theta) \end{aligned} \tag{1.15}$$

Der Wert der posterior pdf bei einem gegebenen Stimuluslevel ergibt sich aus der Wahrscheinlichkeit, dass die bisher gewonnenen Antworten zustande kamen, wenn angenommen wird, dass die Schwelle genau bei dieser Reizstärke liegt.

$$p_{post}(\theta, \mathbf{x}) = \frac{\mathcal{L}(\theta|x_1 \dots x_n)}{\sum_{j=1}^{m} \mathcal{L}(\theta_j|x_1 \dots x_n)} \tag{1.16}$$

Bei der Bayes-Schätzung muss darauf geachtet werden, dass der Bereich der posterior pdf die Schwelle sicher beinhaltet, d.h. dass die Wahrscheinlichkeit der Schwelle am Rand der Verteilung gegen Null gehen sollte. Außerdem muss vermieden werden, dass die posterior pdf am Ende der Messung durch die prior pdf dominiert wird.

Wie man in Gleichung (1.15) sieht, unterscheidet sich die Likelihood-Funktion von der posterior pdf nur durch das Einbeziehen von a priori Informationen mittels Multiplikation mit der prior pdf. Eine Bayes-Schätzung mit dem Maximum der posterior pdf als Schätzwert der Schwelle und einer Rechteckverteilung als prior pdf ist äquivalent zur Maximum Likelihood Schätzung. Die ML-Methode stellt insofern einen Spezialfall der Bayes-Schätzung dar.

Der nächste zu testende Stimulus wird entsprechend des besten aktuellen Schätzwerts der Schwelle gewählt. Bei Maximum Likelihood wird der Stimulus gewählt, der dem Maximum (*mode*) der Likelihood-Funktion entspricht. Andere Prozeduren wie z. B. ZEST nutzen den Median oder auch den Mittelwert (*mean*) der posterior pdf als besten Schätzwert. Nach Pelli (1987) wird eine optimale Stimulusplatzierung dann erreicht, wenn die Likelihood-Funktion schon im Voraus für verschiedene Alternativen berechnet (*look-ahead*) und dann der Stimulus gewählt wird, für den die Varianz der posterior pdf minimal ist [26], [18].
Messungen mit Stimulusplatzierung nach Maximum Likelihood- und Bayes-Prinzip werden meist nach einer vorher festgelegten Anzahl von Trials beendet, obwohl einige Autoren auch ein dynamisches Abbruchkriterium vorsehen. In diesem Fall erfolgt der Abbruch, wenn ein bestimmtes Konfidenzintervall erreicht wird, womit die gewünschte Varianz des Schätzwertes garantiert wäre (nach Laming & Marsh (1988), [10]).

Traditionell wird mit den logarithmierten Werten der Likelihood-Funktion gerechnet um einen *Under-* bzw. *Overflow* der Werte zu verhindern. Da der Logarithmus eine monotone Transformation darstellt, wird der Ort des Maximums dadurch nicht verändert. Die Variante ZEST rechnet jedoch mit nicht modifizierten Wahrscheinlichkeiten, weil die Ermittlung des Mittelwerts der posterior pdf bei transformierten Werten erschwert würde.

Das Problem der parametrischen Verfahren ist die Wahl des zugrunde liegenden psychometrischen Modells, denn die wahre Form und Steigung der psychometrischen Funktion ist meist nicht bekannt. Grobe Schätzungen aus Vorversuchen und Wissen über den Einfluss von Mismatches kann hier sehr hilfreich sein.

Es existieren zahlreiche Varianten parametrischer Verfahren. Zum Beispiel wird bei *Best PEST* (Pentland, 1980) die logistische Funktion zugrunde gelegt und berechnet den Schätzwert nach Maximum Likelihood [19]. QUEST (*Quick Estimation by Sequential Testing*, Watson & Pelli, 1983) und ZEST (*Zippy Estimation by Sequential Testing*, King-Smith et al., 1994) nutzen die Weibull-Form als zugrunde liegendes psychometrisches Modell und führen die Berechnungen nach der Bayes-Methode durch [27], [8].

1.3 Stand der Forschung

In der für den Forschungskontext relevanten Literatur finden sich viele Untersuchungen zu den einzelnen, adaptiven Verfahren und ihren Eigenschaften, wobei sich die Erkenntnisse entweder auf Computer-Simulationen oder auf psychometrische Tests mit Versuchspersonen stützen. Ein Nachteil realer Experimente ist, dass sich die Messungen, die zu einer speziellen Forschungsfrage durchgeführt wurden, nicht unbedingt auf beliebige andere Untersuchungsgegenstände übertragen lassen. Außerdem wird der ermittelte Schwellwert von zufälligen und systematischen Fehlern beeinflusst. In Simulationen können die Verfahren selbst evaluiert und Einflussfaktoren kontrolliert oder gänzlich ausgeschlossen werden. Jedoch muss bei Simulationen ein psychometrisches Modell für die Versuchsperson implementiert werden, wobei nicht sicher ist, ob dadurch das wahre Antwortverhalten einer Versuchsperson bzw. einer Population widergespiegelt wird. Außerdem hat sich gezeigt, dass bei den Antworten realer Versuchspersonen

die Annahmen des stationären Prozesses und der Unabhängigkeit aufeinander folgender Messungen verletzt werden (durch Lapsing, Lerneffekte, etc.). Unklar ist jedoch, wie dieses Verhalten in Computer-Simulationen berücksichtigt werden kann.

1.3.1 Zusammenfassung bisheriger Veröffentlichungen

Kollmeier, Gilkey und Sieben (1988) vergleichen verschiedene adaptive und nichtadaptive Methoden in Simulationen und auch im Experiment mit 4 trainierten Versuchspersonen. Unter den adaptiven Verfahren werden 2Down-1Up und 3Down-1Up Staircase und PEST jeweils mit 2AFC, 3AFC Paradigma kombiniert. Im Vergleich zu den nichtadaptiven Methoden liefern die adaptiven Verfahren etwas bessere Schätzwerte für die Schwelle. Die effizienteste Methode ist dabei die 3Down-1Up Staircase (79% korrekt) mit 3AFC. Die schlechteste Performance zeigt sich bei 2Down-1Up Staircase (71% korrekt) in Verbindung mit 2AFC Paradigma [9].

Die Simulationen von **Schlauch und Rose** (1990) widmen sich speziell dem Vergleich verschiedener Paradigmen und deren Einfluss auf Effizienz und Bias. Dabei kombinieren sie 2-, 3- und 4AFC jeweils mit 2- und 3Down-1Up Staircase. Die Evaluation zeigt, dass die Varianz des Schätzwertes mit zunehmender Intervallzahl und bei höherem Konvergenzniveau sinkt. 3- und 4AFC sind demnach auch bei größerem Zeitaufwand effizienter als 2AFC-Verfahren. Weiterhin wurde ermittelt, das große Schrittweiten zu höherer Varianz und größerem Bias führen [22]. Diese Erkenntnisse decken sich mit den Ergebnissen von Kollmeier et al. (1988).

Shelton, Picardi und Green (1982) untersuchen die Eigenschaften von Adaptive Staircase, Maximum-Likelihood (ML) und PEST-Verfahren mit 2AFC-Paradigma. Die Messungen, die mit drei Versuchspersonen durchgeführt wurden, liefern nicht allzu große Unterschiede. Es zeigt sich jedoch die Tendenz, dass Staircase und ML bei wenigen Trials (< 30) einen leichten Bias der Schätzwerte aufweisen. Diese Abweichungen könnten jedoch durch randomisierte Anfangsstimuli fast gänzlich korrigiert werden. Innerhalb von nur 10 Trials konvergiert nur die ML-Methode, wenn auch mit leichtem

Bias. Dieses Verfahren ist demnach bei Untersuchungen mit Kindern, Patienten oder Tieren empfehlenswert, wenn nur kurze Messzeiten möglich sind. Die ML-Methode ist jedoch auch problematisch, da nur wenige überschwellige Stimuli dargeboten werden, was bei unerfahrenen Versuchspersonen zu Schwierigkeiten führt und ein vorbereitendes Training erforderlich macht [23]. Insgesamt sind die Schlussfolgerungen von Shelton et al. (1982) kritisch zu betrachten, da es fraglich erscheint, ob die Validität der Ergebnisse bei einer so geringen Anzahl von Versuchspersonen gewährleistet ist.

In Simulationen von **Pentland** (1980) werden verschiedene adaptive Methoden getestet um die Vorteile des Best PEST Verfahrens aufzuzeigen. Dabei werden Standard Staircase, PEST, improved PEST und Best PEST mit Ja-Nein-Paradigma evaluiert und anschließend die *setting accuracy* als Standardabweichung der Schätzwerte über der Anzahl benötigter Trials aufgetragen. Den Ergebnissen zufolge benötigt die Best PEST-Prozedur weitaus weniger Trials als die anderen Verfahren um eine bestimmte Genauigkeit zu erreichen. Außerdem wird beschrieben, dass dieses Verfahren frei von Bias und unempfindlich gegenüber Lapsing und Mismatches hinsichtlich des angenommenen Range oder der Form der psychometrischen Funktion sei [19]. Hier muss hinterfragt werden, ob solch generelle Aussagen über das Verfahren möglich sind. Zum einen wird nur das Ja-Nein-Paradigma untersucht, zum anderen werden die Werte der anderen Verfahren anscheinend aus früheren Simulationen übernommen (siehe Taylor & Creelman, 1967 und Findlay, 1978), wobei nicht klar ist, ob die äußeren Bedingungen dabei vergleichbar gewählt wurden.

Eine weitere Untersuchung, die verschiedene Verfahren in Simulationen und realen Experimenten evaluiert, liefern **Madigan und Williams** (1987). Dabei werden PEST, Best PEST und QUEST jeweils mit Ja-Nein und 2AFC-Paradigma kombiniert und die Effizienz anhand der *setting accuracy* (hier als Standardabweichung der Fehler der Schwellwerte) ermittelt. Die Simulationen zeigen, dass Best PEST und QUEST etwa gleiche Genauigkeit liefern, PEST jedoch etwas weniger effizient arbeitet.

Alle drei Methoden zeigen bei Messungen mit 2AFC generell mehr Varianz des Fehlers als bei Ja-Nein-Paradigma, was auf die erhöhte Ratewahrscheinlichkeit zurückzuführen ist [15].
Deshalb werden spezielle Situationen simuliert, in denen eine erhöhte Varianz des Fehlers durch das 2AFC-Paradigma störend wirken kann. Bei der Simulation des Lerneffekts zeigt sich, dass Best PEST und QUEST die Schwelle besonders bei wenigen Trails unterschätzen, jedoch infolge des *threshold shift* meist überschätzen. Der Einfluss von Lapses ist bei Best PEST am gravierendsten, was sich in einer starken Erhöhung der setting accuracy äußert. Außerdem ist eine generelle Tendenz zu rechtseitigem Bias (Überschätzen der Schwelle) erkennbar, obwohl die Beeinträchtigung durch Lapses nicht sehr groß und eventuell in der Praxis nicht relevant ist [15].
Madigan und Williams untersuchen auch den Einfluss von Fehleinschätzungen (*mismatches*) zwischen angenommenen und wahren Eigenschaften der psychometrischen Funktion, einmal bezogen auf den Range der Stimuli und zweitens bezogen auf die Steigung (*Slope*) der psychometrischen Funktion. Bei steigendem Range-Mismatch zeigt Best PEST im Vergleich zu QUEST eine größere Varianz. Außerdem wird die Schwelle zunehmend unterschätzt. Moderate Abweichungen der angenommenen von der wahren Steigung der psychometrischen Funktion haben kaum Auswirkungen auf den Schätzwert, jedoch wird empfohlen die Steigung eher zu überschätzen um den Einfluss auf die Varianz klein zu halten. In realen Experimenten mit 39 Versuchspersonen zeigen sich keine signifikanten Unterschiede in der Varianz der drei Verfahren (mit 2AFC-Paradigma) [15]. Das bestätigt wiederum die Ergebnisse der Untersuchung von Shelton et al. (1982).

Emerson (1986) vergleicht das Maximum Likelihood-Verfahren mit der Bayes-Methode, um speziell das Verhalten in 2AFC-Abfragen zu untersuchen. In der Evaluation wird bei wechselnder realer Schwelle der Mittelwert und die Standardabweichung der Schätzwerte ermittelt und über der Anzahl der Trials aufgetragen. Die ML-Schätzwerte zeigen generell negativen Bias, der auch noch nach 50 Trials bemerkbar ist. Bei der Bayes-Methode ist kein Bias erkennbar. Nur am Anfang der Messreihe zeigt sich eine Tendenz hin zur Mitte.

Die Standardabweichung ist bei ML-Schätzung generell größer als bei der Bayes-Methode, wobei der *worst case* bei einer Unterschätzung der Schwelle auftritt [3].

King-Smith et al. (1994) untersuchen die Eigenschaften von QUEST und verschiedenen Modifikationen mit Ja-Nein- und 2AFC-Paradigma in einer Reihe von Simulationen. Dabei liegt der Schwerpunkt besonders auf der Evaluation verschiedener Methoden der Stimulusplatzierung. Die Ergebnisse zeigen, dass eine Platzierung anhand des Mittelwerts der Likelihood-Funktion (Mean-QUEST) generell zu besseren Schätzwerten führt als Median-QUEST oder Mode-QUEST. Deshalb wird diese Variante, auch ZEST genannt, als neue vorteilhaftere Methode empfohlen. Die höchste Genauigkeit weist die *Minimum Variance* Methode auf, jedoch erfordert diese Methode auch zusätzlichen Rechenaufwand (*look-ahead*).
King-Smith et al. zeigen, dass die Schätzwerte aus Ja-Nein-Abfragen bei allen getesteten Methoden bedeutend besser sind als diejenigen aus 2AFC-Situationen. Außerdem wurde ermittelt, dass die Wahl der prior pdf kaum Auswirkungen auf die Ergebnisse hat, solange eine begründete Schätzung der Anfangsverteilung vorliegt. Auch hier zeigt sich, dass der Effekt durch Slope-Mismatches in ZEST (Mean-QUEST) gegenüber zufälligen Fehlern klein ist [8].

Auch **Treutwein** (1995) führt Simulationen zu Mismatches der Steigung der psychometrischen Funktion durch. Gezeigt werden die Effekte anhand des YAAP-Verfahrens. Dies ist eine Bayes-Methode, die den Mittelwert der Likelihood-Funktion zur verwendet und bei Erreichen eines bestimmten Konfidenzintervalls beendet wird. Den Ergebnissen zufolge führt eine steile psychometrische Funktion zur Fokussierung der Messungen nahe der Schwelle, verursacht aber auch eine größere Anfälligkeit für Fehleinschätzungen der Rate- und Lapsing-Wahrscheinlichkeit. Das hat zur Folge, dass Instabilitäten auftreten und weitab liegende, „schlechte“ Schätzwerte für die Schwelle ermittelt werden. Im Fall einer flachen psychometrischen Funktion verteilen sich die Präsentationen in einem breiteren Bereich um die Schwelle. Dadurch können Mismatches besser toleriert werden, was bei wenig erfahrenen Versuchspersonen durchaus geeigneter erscheint [26].

1.3.2 Kritik an bestehenden Untersuchungen

Die Ergebnisse der bereits veröffentlichten Untersuchungen bilden einen Ansatz um die Vielzahl der psychometrischen Verfahren zu evaluieren. Eine erschöpfende Gesamtevaluation existiert jedoch noch nicht, da die Autoren meist keine repräsentative Auswahl an Verfahren treffen und sogar ganze Gruppen von Verfahren ignorieren. Einerseits ist dies durch die Entwicklung der Rechentechnik bedingt, die heute umfangreichere Simulationen erlaubt. Andererseits gibt es Uneinigkeit über die Methodik der Evaluation. Uneinheitliche Skalierungen und unterschiedliche Randbedingungen machen den Vergleich der Ergebnisse verschiedener Untersuchungen sehr schwer. Andere Evaluationen sind subjektiv gefärbt und unterstützen nur das vom Autor vorgeschlagene Verfahren. Nur selten werden Untersuchungen zum Einfluss sequenzieller Abhängigkeiten durchgeführt [26]. Insgesamt sind die Ergebnisse lückenhaft, keineswegs einheitlich, teilweise sogar widersprüchlich und daher kaum zufrieden stellend.

2 Simulationen

2.1 Vorbereitung

2.1.1 Methode der Simulation

Bei der Evaluation psychometrischer Verfahren mittels empirischer Tests mit Versuchspersonen werden die Ergebnisse oft von intra- und interindividuellen Abweichungen sowie von systematischen Fehlern überlagert. Es ist meist sehr schwierig diese Fehlerquellen von Einflüssen durch das Verfahren selbst zu trennen. Zum Beispiel lässt sich kaum ermitteln, ob ein Verfahren die Schwelle generell über- oder unterschätzt oder ob bestimmte Abweichungen nur durch eine zu große Varianz der Ergebnisse zustande kommt. In Simulationen können diverse Fehlerquellen, die bei empirischen Messungen auftreten, kontrolliert oder auch eliminiert werden. Zum Beispiel ist die wahre psychometrische Funktion der simulierten Versuchsperson und damit auch die Variabilität der wahren Schwelle bekannt. Die Methode der Simulation scheint hier auch deshalb angebracht, weil die Verfahren selbst evaluiert werden sollen und nicht ein spezieller, psychometrischer Untersuchungsgegenstand. Außerdem kann eine große Anzahl von Trials pro Bedingungsvariation garantiert werden, was die zufälligen Fehler minimiert. Für eine vergleichbare Untersuchung mittels empirischer Versuche bräuchte man sehr viele Versuchspersonen und dennoch wäre eine Verallgemeinerung schwierig. Die Simulationen werden hier realisiert, indem die psychometrischen Verfahren in Matlab (Version 7, 2004) unter Mac OS X (Version 10.3.9) implementiert und anschließend evaluiert werden.

2.1.2 Simulation der Versuchsperson

Zur Generierung simulierter Antworten wird ein psychometrisches Modell entworfen und in Matlab implementiert. Das Modell orientiert sich an anderen Evaluationen, die Monte-Carlo-Simulationen durchgeführt haben [1], [3]. Dabei wird davon ausgegangen, dass die Antworten der Versuchspersonen statistisch unabhängig und binomial verteilt sind. Für jede simulierte Versuchsperson wird eine psychometrische Funktion definiert, die die Wahrscheinlichkeit einer korrekten Antwort abhängig von der Reizstärke angibt. Hier wird nur die Lage der psychometrischen Funktion variiert, die Steigung der psychometrischen Funktion bleibt bei jeder Messung gleich. Da der Versuchsleiter meist nur den groben Bereich kennt, in dem sich die wahre Schwelle befindet, wird bei der Variation der wahren Schwelle systematisch vorgegangen. So wird garantiert, dass jeder Wert aus dem vorbestimmten Bereich mit gleicher Häufigkeit auftritt. In jeder simulierten Abfrage entscheidet dann eine mit der jeweiligen Antwortwahrscheinlichkeit gewichtete Zufallszahl darüber, ob eine korrekte oder falsche Antwort gegeben wird. Die folgende Abbildung 2.1 zeigt beispielhaft das Antwortverhalten einer simulierten Versuchsperson aufgrund eines psychometrischen Modells mit logistischer Form. Die genaue Implementierung der simulierten Versuchsperson ist im Quellcode der simulierten Verfahren zu sehen, der im Anhang B aufgeführt ist.

2.1.3 Auswahl der Paradigmen

In den Simulationen wird stets die Ja-Nein-Abfrage evaluiert, obwohl diese Vorgehensweise in empirischen Messungen zu ungenauen Ergebnissen führt, da sie mit dem individuell verschiedenen Entscheidungskriterium konfundiert sind. Die Ja-Nein-Abfrage kann hier jedoch als Referenz dienen, da in Simulationen kein Kriterium auftritt. Die Verfahren werden weiterhin mit den oft verwendeten Alternative Forced-Choice-Paradigmen kombiniert. Obwohl viele Evaluationen die schlechte Eignung des 2AFC-Paradigma bestätigen, wird es immer noch sehr häufig eingesetzt. Es wird hier in die Evaluation mit einbezogen, um einen Vergleich mit den Ergebnissen von Madigan und Williams (1987) zu ermöglichen.

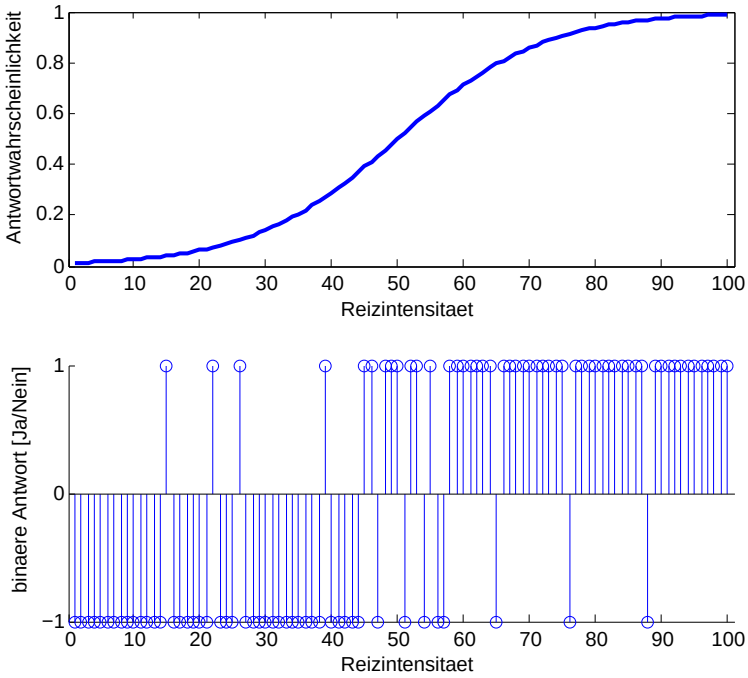

Abbildung 2.1: Simulation binärer Antworten

Das 3AFC-Paradigma gilt wegen der niedrigeren Ratewahrscheinlichkeit als besser geeignet. Es wurde jedoch noch nicht mit vielen Verfahren getestet und wird daher auch in die Evaluation miteinbezogen.

2.1.4 Auswahl der Verfahren

Eine komplette vergleichende Betrachtung der existierenden, adaptiven Verfahren wäre auch hier zu umfangreich. Zu viele Varianten psychometrischer Prozeduren müssten unter verschiedensten Bedingungsvariationen getestet werden. Jedoch erscheint eine systematische Untersuchung einer repräsentativen Auswahl psychometrischer Verfahren zur Ergänzung und Erweiterung schon bestehender Evaluationen durchaus angebracht. Dabei wird aus jeder der oben genannten Verfahrens-Gruppen ein Vertreter gewählt, um die verschiedenen Herangehensweisen einzubeziehen. Somit werden nicht nur aufwendige, parametrische Verfahren getestet, sondern auch einfache Algorithmen, die kaum Annahmen über die zugrunde liegende psychometrische Funktion machen.

Bei der Wahl der Parameter wird ähnlich vorgegangen wie bei den Simulationen von Madigan und Williams (1987) um einen Vergleich der Ergebnisse zu ermöglichen. Die Auswahl der Verfahren und der Paradigmen wird bezogen auf diese Untersuchung jedoch noch erweitert. Herausforderung ist es, die verschiedenen, wenn auch nur leichten, Vorzüge der Prozeduren zu ermitteln, die in bestimmten Versuchskonstellationen von Vorteil sein könnten.

Auswahl aus Staircase-Verfahren

Das Simple Staircase-Verfahren als eines der ältesten, adaptiven Verfahren wird immer noch oft verwendet und soll deshalb hier als Vertreter der Staircase-Methoden gewählt werden. Der Nachteil dieses Algorithmus ist jedoch, dass er immer bei 50% korrekt konvergiert. Bei der Kombination mit 2AFC- und 3AFC-Paradigma würde sich ein ratekorrigiertes Konvergenzniveau von jeweils 0% und 25% korrekt ergeben. Dies stellt jedoch eine schlechte Voraussetzung dar, da kein Pendeln um die Schwelle möglich wäre. Durch die Anpassung der adaptiven Regel nach Levitt (1971) können auch andere Konvergenzniveaus erreicht werden. Deshalb wird für die AFC-Paradigmen hier die Transformed Up-Down-Methode gewählt. Um die beste Annäherung an das 50%-Niveau zu erreichen, wird hier das 2AFC-Paradigma mit 3Down-1Up Staircase kombiniert, das 3AFC-Paradigma jedoch mit der 2Down-1Up-Regel.

PEST versus More Virulent PEST

Das PEST-Verfahren hat sich als ein sehr stabiles Verfahren etabliert. Die 1978 von Findlay vorgeschlagenen Modifikationen sollen es durch ein variables Deviation Limit W noch schneller konvergieren lassen. Vorbereitende Simulationen zeigen jedoch, dass das improved PEST keine besseren Ergebnisse liefert. Zwar führen die Modifikationen zu einer höheren Anzahl von effektiven Reversals bei fester Anzahl von Trials, jedoch sind die Ergebnisse nicht genauer. Im Gegenteil, PEST weist weniger Varianz der Fehler und auch weniger Bias auf als die vermeintlich verbesserte Prozedur von Findlay. Folglich wird hier die ältere PEST-Methode implementiert.

Auswahl aus Maximum Likelihood-Verfahren

Als Vertreter der Maximum Likelihood-Verfahren kommen Hall's Hybrid-Methode (Hall, 1968), Best PEST (Pentland,1980) und ML-Test (Harvey, 1986) in Frage [26]. Die Prozedur von Hall ist allerdings darauf ausgelegt neben der Lage auch die Steigung der psychometrischen Funktion zu schätzen. Die beiden anderen Varianten ermitteln nur die Lage der Schwelle, ML-Test nutzt jedoch ein dynamisches Abbruchkriterium, was hier nicht der Fall sein soll. Best PEST wird von Pentland als sehr effiziente und von Bias freie Prozedur gelobt [19], [12]. In vorbereitenden Simulationen mit AFC-Paradigma weist Best PEST jedoch überraschend schlechte Werte auf. Da diese Methode in empirischen Tests häufig mit 3AFC-Paradigma verwendet wird, scheint es angebracht die bisherigen Ergebnisse zu Best PEST zu überprüfen und die Erkenntnisse bezogen auf verschiedene Paradigmen zu erweitern.

Auswahl aus Bayes-Methoden

Hier kommen auch verschiedene sehr ähnliche Methoden in Frage: QUEST (Watson & Pelli,1983), Quadrature Method (Emerson, 1986), IDEAL (1987, Pelli), YAAP (Treutwein, 1989) und ZEST (King-Smith et al., 1991) [26]. Die QUEST-Methode wurde schon bei Madigan & Williams und auch bei King-Smith evaluiert [15], [8]. Auch eigene, vorbereitende Simulationen bestätigen, dass QUEST ähnliche Eigenschaften aufweist wie Best PEST, da zur Platzierung der Stimuli auch das Maximum der Wahrscheinlichkeitsverteilung (posterior pdf) verwendet wird. King-Smith et al. haben gezeigt, dass eine Stimulusplatzierung im Mittelwert der Wahrscheinlichkeitsverteilung die besten Ergebnisse liefert und schlagen deshalb ZEST als genauere und effizientere Methode vor. Da ZEST (im Gegensatz zu YAAP) nach fester Anzahl von Trials endet, wird diese Methode hier als Vertreter der Bayes-Schätzung für die Simulation ausgewählt.

Damit ergibt sich die folgende Auswahl an Verfahren:

- Simple Staircase bzw. Transformed Up-Down (je nach Paradigma)
- PEST
- Best PEST
- ZEST

Jede der vier Methoden wird jeweils mit den drei Paradigmen (Ja-Nein, 2AFC und 3AFC) kombiniert, wodurch sich 12 verschiedene Varianten ergeben.

2.2 Implementierung der Verfahren

2.2.1 Allgemeines

Stimulus Set

Um die Vergleichbarkeit der Ergebnisse zu gewährleisten, wird für alle Simulationen eine einheitliche Skalierung in *logit units* gewählt. Eine logit unit wird definiert als der Betrag der Steigung im Wendepunkt der logistischen Kurve, die die wahre psychometrische Funktion der Versuchsperson beschreibt. Demzufolge wird ein Reiz, der 1 logit unit oberhalb der Schwelle liegt, in einer Ja-Nein-Abfrage mit einer Wahrscheinlichkeit von 0.73 erkannt. Zwischen einem Reiz mit 50%-Korrektwert und einem Reiz mit 99%-Korrektwert liegt ein Abstand von 4.595 logit units. Für die Simulationen liegen die Stimuli im Bereich von -5.0 bis 5.0 logit units in Stufen von je 0.25 logit units vor. Es ergibt sich ein Set von 41 diskreten, gleichabständigen Stimuli.

Psychometrisches Modell der Versuchsperson

Für das psychometrische Modell der simulierten Versuchsperson wird die logistische Funktion gewählt. Die beiden Freiheitsgrade sind der Threshold- und der Slope-Parameter, die jeweils die Lage und die Steigung der Funktion beeinflussen. Bei höherer Steigung konvergiert die Prozedur schneller gegen den Schwellwert, jedoch wird sie dadurch

auch störanfälliger gegenüber Lapses. Flachere psychometrische Funktionen bewirken, dass die Stimuli in einem größeren Bereich um die Schwelle platziert werden, was eine langsamere Konvergenz zur Folge hat. Die Wahl der Parameter orientiert sich im Folgenden an den Simulationen, die von Madigan & Williams durchgeführt wurden. Es wird ein mittlerer Slope von 1 logit unit gewählt. Dies entspricht einem Verhältnis von Range/10. Bei jeder neuen Messreihe (*Run*) wird ein anderer wahrer Schwellwert vorgegeben. Die wahre Schwelle variiert im Bereich ± 3 logit units um die Mitte des Range. Da die wahre Schwelle in empirischen Versuchen vor der Messung nicht bekannt ist, werden die wahren Schwellwerte in den Simulationen systematisch variiert, um eine Gleichverteilung zu erzeugen. Der Mittelwert aller wahren Schwellwerte liegt bei 0 logit units, die Standardabweichung bei 1,8 logit units (nach 1000 Runs). Je nach dem verwendeten Paradigma muss auch das Antwortverhalten der Versuchspersonen angepasst werden. Bei 2AFC liegt die Ratewahrscheinlichkeit bei 50% und die untere Asymptote der psychometrischen Funktion daher bei $p = 0.5$. Bei 3AFC beträgt die Ratewahrscheinlichkeit 33.3% und die untere Asymptote liegt bei $p = 1/3$). Bei Ja-Nein Paradigma wird die Ratewahrscheinlichkeit auf 3% gesetzt. Die Rate falscher Zuweisungen im überschwelligen Bereich (Lapsing Rate) wird wie auch bei King-Smith [8] auf 2% angesetzt. Die psychometrische Funktion nähert sich also einer oberen Asymptote bei $p = 0.98$ an.

Initial Stimulus Level

Um bei der Wahl des ersten dargebotenen Stimulus keine Reizstärke zu bevorzugen, wird der Wert des Anfangsstimulus meist in einem bestimmten Bereich um die vermutete Schwelle randomisiert. Allerdings nutzen einige der hier getesteten Verfahren a priori Informationen um den ersten Stimulus möglichst dort zu platzieren, wo die Schwelle vermutet wird. Damit diesen Verfahren kein Vorteil eingeräumt wird, beginnen die Simulationen für alle Verfahren immer in der Mitte des Range (bei $x = 0$ logit units). Vorbereitende Simulationen zeigen auch, dass sich die Ergebnisse bei festen und randomisierten Anfangsstimuli nicht merklich unterscheiden.

Abbruchkriterium

Meist wird ein Abbruchkriterium nach fester Anzahl von Trials verwendet, wodurch die Vergleichbarkeit verschiedener Verfahren bezogen auf Genauigkeit und Effizienz erst ermöglicht wird. Die 12 Verfahren werden systematisch über eine vorbestimmte Anzahl von 10, 20, 30, 40 und 50 Trials (Abfragen pro Schätzwert) getestet. Jede der 60 Testvarianten wird in jeweils 1000 Runs evaluiert (Wiederholung der Messung mit 1000 Versuchspersonen).

Im folgenden Abschnitt wird auf die Implementierung der einzelnen Verfahren eingegangen. Eine Übersicht zu den gewählten Verfahren und Parametern findet sich in Tabelle 2.2. Der ausführliche Quellcode der implementierten Verfahren ist im Anhang unter Abschnitt B einzusehen.

2.2.2 Implementierung des Staircase-Verfahrens

Die Simple Staircase Methode (1Down-1Up) konvergiert bei Ja-Nein-Paradigma bei $p = 0.5$. Nach einer positiven Antwort wird die Reizstärke gesenkt, nach einer negativen Antwort wird sie erhöht. Für das 2AFC-Paradigma wird die 3Down-1Up-Regel implementiert. Dabei wird die Reizstärke erst nach drei aufeinander folgenden richtigen Zuweisungen gesenkt. Man erhält ein Konvergenzniveau von $p = 0.794$, dass sich durch die Korrektur nach Abbott (Gl. 1.3) bei 2AFC auf $p = 0.61$ verringert. Für das 3AFC-Paradigma ergibt sich in Kombination mit 2Down-1Up ein ratekorrigiertes Konvergenzniveau von $p = 0.58$. Da das 50%-Niveau nicht ganz erreicht wird, ist bei den Simulationen ein geringer positiver Bias zu erwarten. Die Werte stellen jedoch die beste Annäherung an das gewünschte Konvergenzniveau dar. Abgesehen von der Stimulus-Platzierung, ändert sich bei dieser Anpassung kaum etwas an den Verfahren. Anfangsstimulus, Schrittweite und Abbruchkriterium sind bei allen drei Varianten gleich.

Übersicht zur Simulation adaptiver, psychometrischer Verfahren

	Staircase			PEST (Taylor & Creelman 1967)	Best PEST (Pentland 1980)	ZEST (King-Smith et al. 1994)
Eigenschaften	Anpassung der Stimulusstärke Eine Halbierung der Schrittweite nach erstem Reversal			adaptive Schrittweite u. Reizstärke SPRT, Wald Test (Parameter W) Platzierung nach heuristischen Regeln	adaptive Schrittweite u. Reizstärke Maximum Likelihood Schätzung a priori Infos durch implizite Trials	adaptive Schrittweite u. Reizstärke Bayes-Schätzung (Mean) prior density function
Paradigma	Yes-No	2AFC	3AFC	Yes-No / 2AFC / 3AFC	Yes-No / 2AFC / 3AFC	Yes-No / 2AFC / 3AFC
Ratewahrscheinlichkeit	0.03	0.5	0.33	0.03 / 0.5 / 0.33	0.03 / 0.5 / 0.33	0.03 / 0.5 / 0.33
Adaptive Regel	Simple Up-Down	3Down-1Up	2Down-1Up	-	-	-
Konvergenzniveau (korrigiert)	0.495	0.612	0.578	0.505 / 0.74 / 0.656	0.505 / 0.74 / 0.656	0.505 / 0.74 / 0.656
Anzahl der Messungen	1000	1000	1000	1000	1000	1000
Anzahl der Trails	10	10	10	10	10	10
	20	20	20	20	20	20
	30	30	30	30	30	30
	40	40	40	40	40	40
	50	50	50	50	50	50
Range [logit units]	-5 ... 5	-5 ... 5	-5 ... 5	-5 ... 5	-5 ... 5	-5 ... 5
Auflösung [logit units]	0.25	0.25	0.25	0.25	0.25	0.25
Reale psychometr. Funktion	logistisch	logistisch	logistisch	logistisch	logistisch	logistisch
Reale Steigung [logit units]	1	1	1	1	1	1
Reale Lapsing Rate	0.02	0.02	0.02	0.02	0.02	0.02
Reale Schwelle [logit units] **(systematisch variiert)**	-3 ... 3	-3 ... 3	-3 ... 3	-3 ... 3	-3 ... 3	-3 ... 3
psychometr. Funktion	-	-	-	-	logistisch	logistisch
Steigung [logit units]	-	-	-	-	1	1
Lapsing Rate	-	-	-	-	0.02	0.02
Initial Stimulus [logit units]	0	0	0	0	0	Mittelwert der prior density (0)
Initial Step Size [logit units]	2	2	2	4	-	-
Parameter W	-	-	-	1.5 / 1 / 1	-	-
Final Step Size [logit units]	1	1	1	0.25	0.25	0.25
Final Estimate	Mittelwert der Turnarounds			letztes berechnetes Stimuluslevel	Maximum der Likelihood-Funktion letzter ermittelter Teststimulus	Mittelwert der posterior pdf letzter ermittelter Teststimulus

Abbildung 2.2: Übersicht zu gewählten Verfahren und Parametern

Bei Staircase-Verfahren sind Effizienz und Genauigkeit stark abhängig von der Wahl der Schrittgröße, da diese während der Messung nicht variiert wird. Als Modifikation schlagen einige Autoren vor, die Schrittweite nach einer bestimmten Anzahl von Reversals zu halbieren. Dies wurde hier auch implementiert. In vorbereitenden Simulationen wurde getestet, welche Anfangsschrittweite zu optimalen Ergebnissen führt. Die Schrittweite wird auf anfangs 2 logit units gesetzt und nach dem ersten Reversal auf 1 logit unit halbiert. Abbildung 2.3 zeigt den typischen Verlauf einer Messung mit 2Down-1Up-Staircase.

Die Messung endet nach der vorgegebenen Anzahl von Trials und der resultierende Schätzwert ergibt sich als Mittelwert der Reizstärken an den gemessenen Umkehrpunkten, wobei das erste Reversal nicht berücksichtigt wird. Bei kleiner Anzahl von Trials ergibt sich allerdings häufig der Fall, dass insgesamt nur eine Umkehrung stattfindet. Bei diesen Messreihen muss somit die erste Umkehrung in die Berechnung des Schätzwerts miteinbezogen werden, um die Mittelwertbildung zu ermöglichen. Diese Anpassung könnte zu etwas größeren Varianzen bei den betroffenen Messreihen führen. Im Allgemeinen wird bei der Staircase-Prozedur erwartet, dass sie zwar keine sehr genauen, aber von Bias freie Schätzwerte liefert.

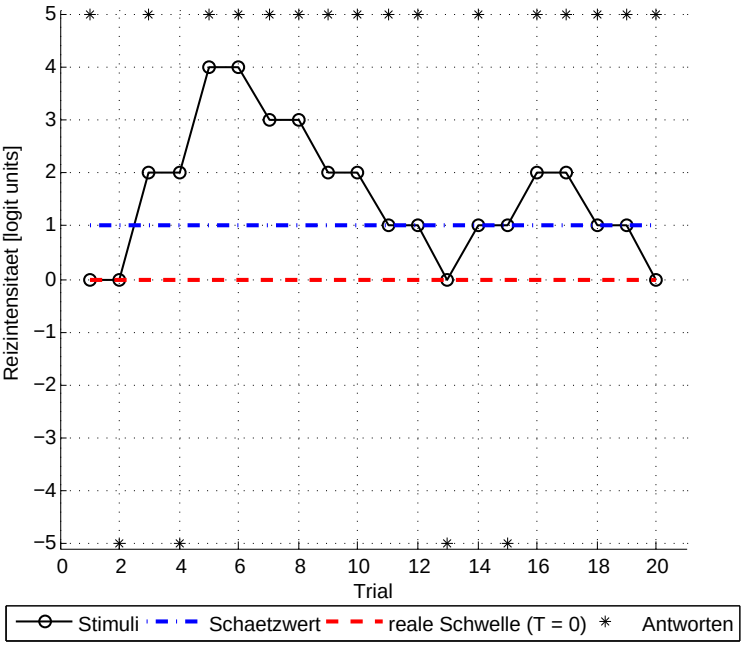

Abbildung 2.3: Verlauf einer Messung mit 2Down-1Up Staircase

2.2.3 Implementierung von PEST

Wie in 1.2.6 beschrieben, wird das PEST-Verfahren mit dem vereinfachten SPRT-Test (Wald, 1947) implementiert, um zu steuern, wann die Reizstärke verändert werden muss. Die Schrittweitenanpassung erfolgt auch in den Simulationen nach den von Taylor & Creelman vorgeschlagenen, heuristischen Regeln. Zu Beginn der Testreihe muss die anfängliche Schrittgröße festgelegt werden, die sich dann im Laufe der Messung je nach Anzahl der Trials immer weiter halbiert. Außerdem muss das deviation limit W gewählt werden. Eigene Simulationen zeigen, dass ein Wert von $W = 1$ bei erhöhter Ratewahrscheinlichkeit die besten Ergebnisse liefert. Dieser Wert wird für die Anwendung bei 2AFC und 3AFC-Paradigma genutzt. Für das Ja-Nein-Paradigma wurde ein optimaler Wert von $W = 1.5$ ermittelt. Das Konvergenzniveau ϕ ist frei wählbar und ändert sich hier je nach Paradigma.

$$\phi = (1 - p_l + p_g)/2 \tag{2.1}$$

Abbildung 2.4 zeigt den typischen Verlauf der PEST-Prozedur. Die Messung endet nach Vorgabe der Autoren automatisch, wenn eine bestimmte Schrittgröße unterschritten wird. Die zuletzt getestete Reizstärke ist zugleich die resultierende Schwelle. Hier muss die Implementierung jedoch so angepasst werden, dass die Messung nach einer festen Anzahl von Trials endet. Dabei muss beachtet werden, dass die kleinst mögliche Auflösung der Stimuli (0.25 logit units) nicht unterschritten werden darf. Wird die Messung beendet bevor die vorgegebene kleinste Schrittgröße erreicht ist, hat diese Anpassung keinerlei beeinträchtigende Wirkung auf die Ergebnisse. Ein Problem ergibt sich jedoch, wenn die Messung weiterläuft, obwohl die kleinstmögliche Schrittgröße schon erreicht ist. Die Schätzung der Schwelle kann dann auch bei größerer Anzahl von Trials nicht mehr genauer werden. Deshalb wurden vorbereitende Simulationen durchgeführt um die anfängliche Schrittweite an die durchschnittliche Anzahl von effektiven Halbierungen anzupassen. Bei 50 Trials ergibt sich eine Mittelwert von 4.5 effektiven Reversals. Daher wurde eine Anfangsschrittweite von 4 logit units und eine minimale Schrittweite von 0.25 logit units für die Simulationen gewählt. Bei PEST ist zu erwarten, dass diese Methode genauere Schätzwerte liefert als das Staircase-Verfahren. Allerdings wird PEST kaum bessere Ergebnisse liefern als die parametrischen Verfahren Best PEST und ZEST, da diese Prozedur oft weitaus mehr Trials benötigt um zu konvergieren.

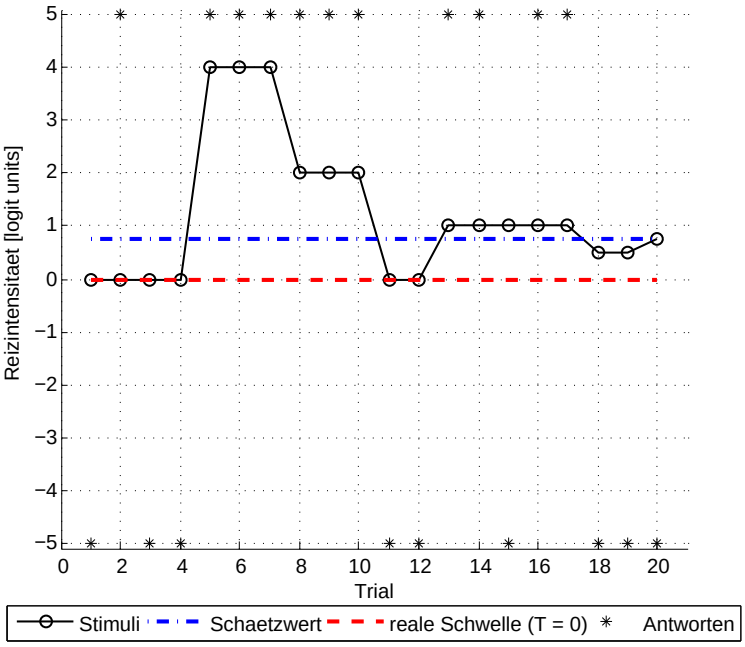

Abbildung 2.4: Verlauf einer Messung mit PEST-Methode

2.2.4 Implementierung von Best PEST

Wie schon in 1.2.6 erwähnt, erfolgt die Stimulusplatzierung bei Best PEST nach Maximum Likelihood-Schätzung (kurz ML-Schätzung). Dazu müssen Form, Range und Steigung der psychometrischen Funktion vorher möglichst genau geschätzt werden. Nach Pentland wird die logistische Funktion verwendet (Gl. 2.2), wobei der Range mit ± 5 logit units, der Slope-Parameter bei $\beta = 1$ logit unit und die Lapsing Rate bei $p_l = 2\%$ gewählt wird. Diese Werte entsprechen also denjenigen der Versuchsperson. Im Vergleich zu Ja-Nein-Abfragen erhöht sich bei Forced-Choice-Abfragen die Ratewahrscheinlichkeit auf Seite der simulierten Versuchsperson. Das heißt, dass auch auf Seite der zugrunde liegenden psychometrischen Funktion eine Anpassung der Parameter nötig ist.

$$\phi(x) = p_g + (1 - p_g - p_l)\frac{1}{(1 + e^{-\beta(x-\theta)})} \tag{2.2}$$

Vor Beginn der Messung werden implizite Trials implementiert, die zur Vorformung der Likelihood-Funktion führen. Dies soll zu große Reizstärkenunterschiede vermeiden und eine bessere Platzierung der Stimuli besonders am Anfang der Messung bewirken.

Bei Ja-Nein-Paradigma wird die größte Reizstärke einmal als erkannt und die kleinste Reizstärke einmal als nicht erkannt vorgegeben. Der erste dargebotene Stimulus liegt damit in der Mitte des Range. Bei 2AFC wird die kleinste Reizstärke einmal als richtig erkannt und einmal als nicht erkannt definiert. Dies soll die Ratewahrscheinlichkeit von 50% berücksichtigen. Bei 3AFC wird analog die Ratewahrscheinlichkeit von 33.3% mittels drei impliziter Trials vordefiniert. Bei den AFC-Verfahren wird der erste Stimulus wie auch beim Ja-Nein-Paradigma in der Mitte des Range platziert. Basierend auf allen vorhandenen Daten werden nach jeder Antwort die Wahrscheinlichkeiten für jeden möglichen Stimulus aktualisiert und der Reiz berechnet, der mit größter Wahrscheinlichkeit auf die gesuchte Schwelle führt. Dieser beste Schätzwert wird dann als nächstes dargeboten. In Abbildung 2.5 sind die Likelihood-Kurven der ersten fünf Trials einer Messung mit Maximum Likelihood dargestellt. Dabei zeigen die beiden flacheren Kurven das Ergebnis der impliziten Trials.

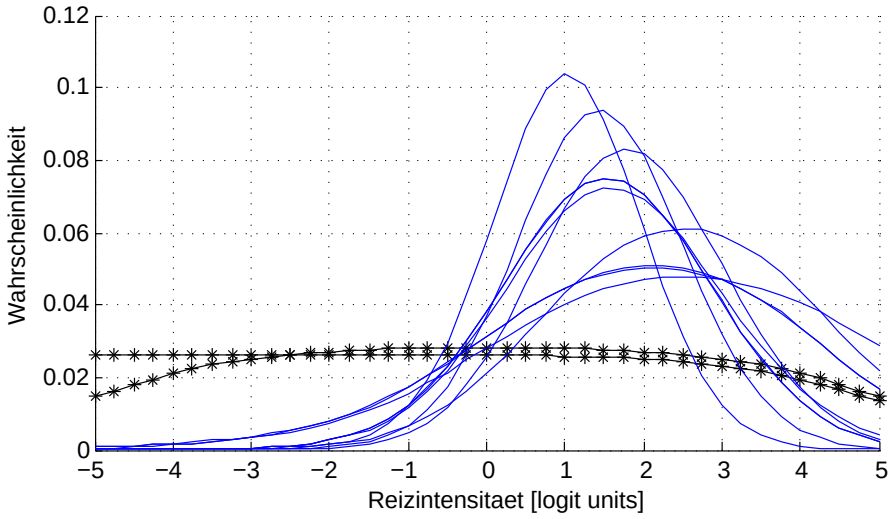

Abbildung 2.5: Likelihood-Kurven bei Best PEST nach 5 Trials

Die von Pentland implementierte Logarithmierung der Wahrscheinlichkeiten zur Vermeidung zu kleiner Werte (Underflow) ist heute nicht mehr nötig. Die Messung endet nach vorgegebener Anzahl von Trials. Die resultierende Schwelle ergibt sich aus dem zuletzt ermittelten ML-Schätzwert. Abbildung 2.6 zeigt den typischen Verlauf einer Messung mit Best PEST und Ja-Nein-Paradigma.

Für die Simulationen wird erwartet, dass Best PEST genauere Schätzwerte ermittelt als Staircase und PEST. Vorbereitende Simulationen lassen vermuten, dass dieses Verfahren in Kombination mit AFC-Paradigma eine erhöhte Varianz und negativen Bias aufweist.

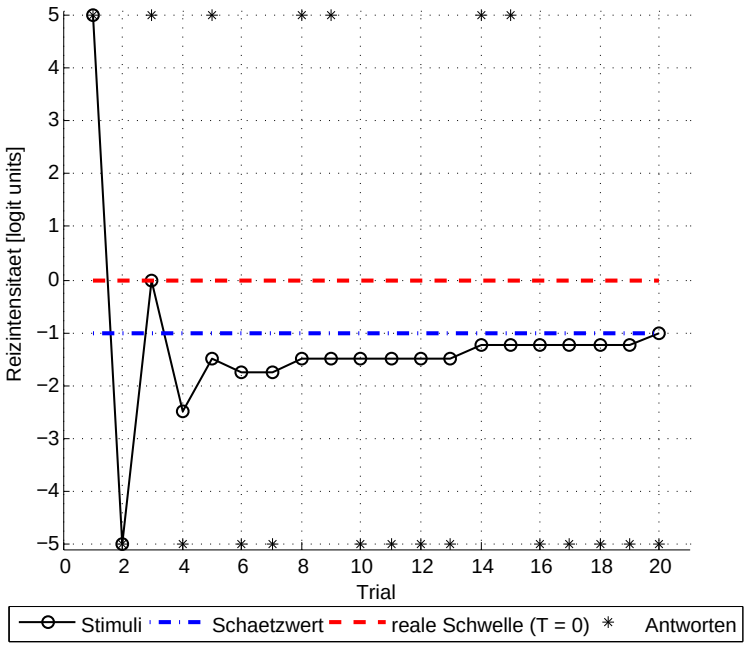

Abbildung 2.6: Verlauf einer Messung mit Best PEST-Methode

2.2.5 Implementierung von ZEST

Das parametrische Verfahren ZEST basiert auf der Bayes-Schätzung, wobei die Autoren dieses Verfahrens die Weibull-Funktion als psychometrisches Modell zugrunde legen. Voraussetzung für die Verwendung der kumulierten Weibull-Verteilung ist jedoch, dass die physikalischen Reize vor der Messung logarithmisch skaliert werden. In der Messung wird dann mit den gleichabständigen Reizen x der Transformationsebene gerechnet.

$$\text{logarithmische Skalierung:} \quad x = c \log_{10}(S) + k \tag{2.3}$$

$$\text{Weibull-Funktion:} \quad \phi(x) = p_g p_l + (1 - p_l)[1 - (1 - p_g)e^{-10^{\beta(x-k)/c}}] \tag{2.4}$$

Die logistische und die Weibull-Form ähneln sich zwar, nehmen aber nie ganz identische Werte an. Ein Mismatch zwischen wahrer und angenommener psychometrischer Form würde also zu ungenaueren Werten führen. Zum Vergleich sind die beiden Kurven in Abbildung 2.7 dargestellt, wobei die Freiheitsgrade der Weibull-Funktion so angepasst wurden, dass ein möglichst ähnlicher Verlauf wie bei der logistischen Funktion entsteht ($\beta = 3.2$, $k = 0$, $c = 1$).

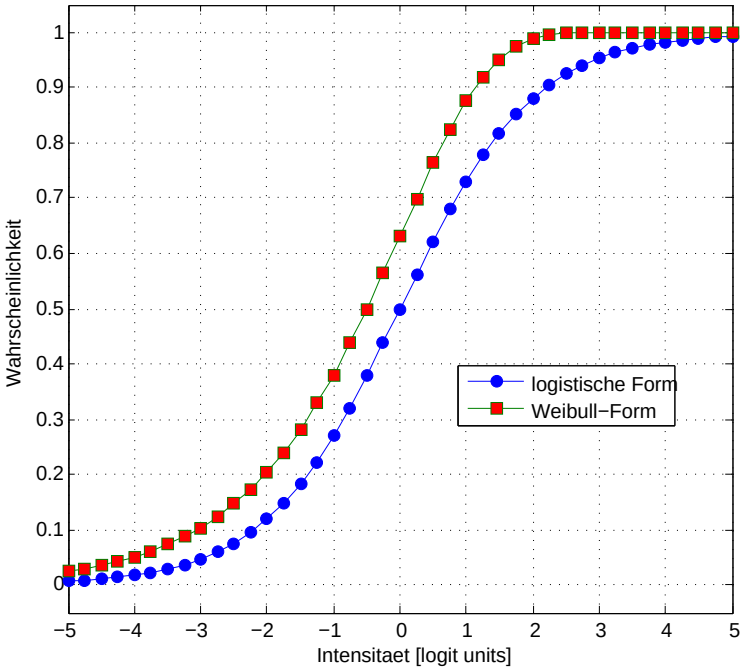

Abbildung 2.7: Weibull und logistische Funktion im Vergleich

Um dem Best PEST-Verfahren keinen Vorteil gegenüber ZEST zu verschaffen, wird letzteres mit der logistischen Funktion implementiert. Da die Wahl des zugrunde liegenden psychometrischen Modells unabhängig von der Wahl des Verfahrens ist, kann die psychometrische Funktion ohne Weiteres gewechselt werden. Die ZEST-Prozedur in Verbindung mit der logistischen Funktion wurde außerdem auch schon bei anderen Evaluationen verwendet [20]. Analog zur Implementierung von Best PEST wird für den Slope-Parameter $\beta = 1$ logit unit gewählt. Auch die Lapsing Rate p_l wird für die zugrunde liegende psychometrische Funktion wieder auf 2% gesetzt. Die Ratewahrscheinlichkeit p_g verhält sich analog und wird auf das jeweilige Paradigma angepasst.

ZEST nutzt a priori Informationen über die wahrscheinliche Lage der Schwelle, die hier als Gauß-förmige prior pdf $g(\theta)$ in die Schätzung miteinbezogen werden. In Abbildung 2.8 sind die Kurven der prior und posterior pdf geplottet, die sich nach den ersten 10 Trials einer Messung mit ZEST ergeben. Durch die hier verwendete prior pdf ist nur eine leichte Vorformung der Wahrscheinlichkeitsverteilung gegeben. ZEST nutzt demzufolge gleiche Voraussetzungen wie Best PEST. Die Messreihe endet nach der vorbestimmten Anzahl von Trials. Die resultierende Schwelle ergibt sich als Mittelwert der zuletzt berechneten posterior pdf. Abbildung 2.9 zeigt den typischen Verlauf einer Messung mit ZEST und Ja-Nein-Paradigma. Der Literatur zufolge sollte ZEST noch genauere Ergebnisse liefern als Best PEST. Außerdem ist für die Simulation zu erwarten, dass die Schätzwerte weniger durch Bias beeinflusst werden.

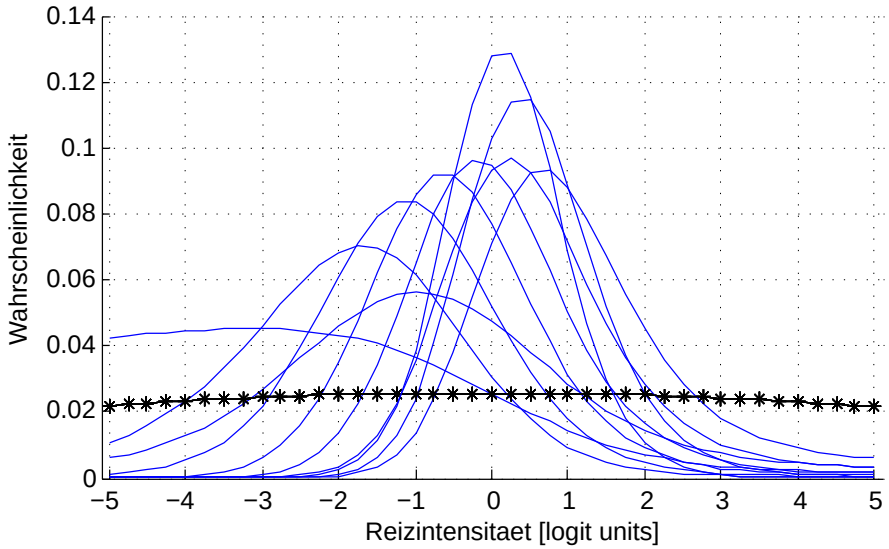

Abbildung 2.8: Prior und Posterior pdf bei ZEST nach 10 Trials

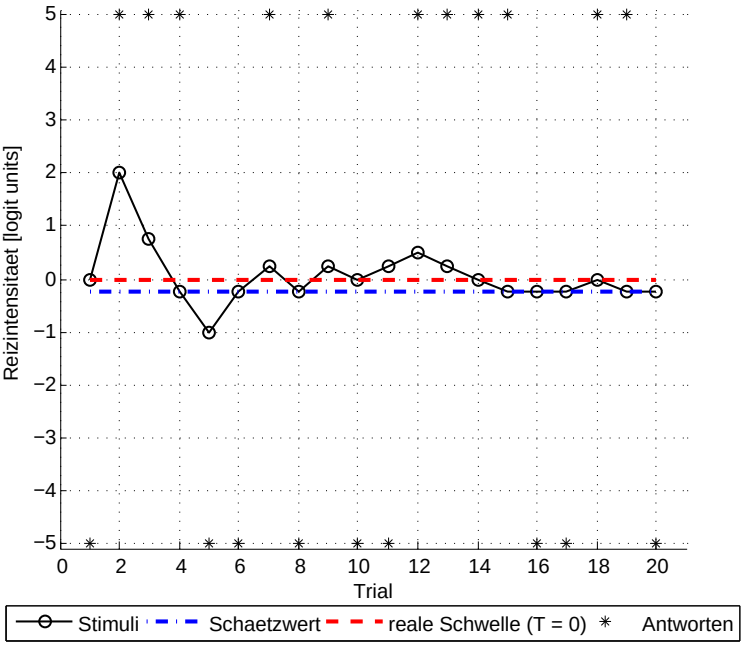

Abbildung 2.9: Verlauf einer Messung mit ZEST-Methode

2.3 Ergebnisse

Die Kombination der 4 psychometrischen Methoden mit 3 Paradigmen und 5 verschiedenen Triallängen ergibt 60 Messreihen, die jeweils 1000 ermittelte Schätzwerte umfassen. Für eine Messreihe werden alle vorher festgelegten Parameter, die Informationen zum Verlauf der Messungen (dargebotene Stimuli und korrespondierende Antworten) und die ermittelten Daten zur wahren und geschätzten Schwelle in einem mat-file gespeichert (siehe Ordner *Simulationsdaten* auf der beiliegenden CD).

Zur Betrachtung der Ergebnisse, gruppiert nach Verfahren, werden die Messwerte im folgenden Abschnitt zuerst in Boxplots dargestellt. Diese zeigen auf einen Blick die Verteilung bzw. Konzentration der Werte an. Neben dem Betrag der Streuung werden auch Asymmetrien und etwaiger Bias sichtbar. Dargestellt werden der Median (als waagerechte rote Linie), der Bereich der Quartile (als blaue Box), der Umfang des 95%-Konfidenzintervalls (als Whisker, schwarze Strich-Punkt-Linie) und Ausreißer (als rote Kreuze). Die Verteilungen werden jeweils über den Bedingungen Trialzahl und Paradigma aufgetragen. Hier wird zuerst die Verteilung der ermittelten Schätzwerte und dann die Streuung der Fehler in jeweils vier Boxplots betrachtet. Der Fehler der Schwellenschätzung wird jeweils als Differenz zwischen ermitteltem und wahrem Schwellwert bestimmt. Der Vorteil dabei ist, dass die Variation der wahren Schwellwerte dadurch herausgerechnet wird und verfahrenseigene Unterschiede deutlicher sichtbar werden.

In Abschnitt 2.3.2 werden die Verfahren dann direkt miteinander verglichen. Dazu werden die in Abschnitt 1.2.4 beschriebenen Bewertungsgrößen zu Varianz, Bias und Effizienz für jede der 60 Messreihen berechnet und in jeweils einem Diagramm über der Anzahl der Trials aufgetragen.

2.3.1 Ergebnisse zu den einzelnen Verfahren

Staircase

Abbildung 2.10 zeigt die Verteilung der Schätzwerte beim Staircase-Verfahren. Die genauen Werte der Perzentile sind im Anhang in Tabelle A.1 aufgeführt.

Zuerst wird die Symmetrie der Schätzwert-Verteilungen betrachtet. Bei Staircase mit Ja-Nein-Paradigma liegen Median und Perzentile immer symmetrisch um 0 logit units verteilt. Bei der 3Down-1Up-Methode und 2AFC zeigt sich bei 10 Trials ein positiver Bias von etwa 1 logit unit. Mit zunehmender Anzahl von Trials reduziert sich diese Abweichung auf bis zu 0.25 logit units (bei 50 Trials). Der Median liegt bei 2AFC nicht immer in der Mitte der Verteilung. Auch bei 2Down-1Up Staircase und 3AFC ist positiver Bias vorhanden, der jedoch im Vergleich zum 2AFC-Paradigma etwas kleiner ausfällt (0.5 bis 0.25 logit units). Um den Betrag der Streuung zu beurteilen, werden die Inter-Perzentil-Abstände betrachtet. Generell ist zu erkennen, dass sich die Streuung mit zunehmender Anzahl von Trials verringert. Diese Entwicklung sieht man am deutlichsten an der zunehmenden Verkürzung der Whisker, die jeweils das 2.5- und 97.5%-Perzentil darstellen. Der Interquartil-Abstand verändert sich bei Ja-Nein-Paradigma kaum (3logit units) und bei den Messungen mit AFC-Paradigma nur gering (von anfangs 3.75 auf 3 logit units). Die Länge der Whisker verringert sich jedoch deutlich. Zum Beispiel bei 2AFC von 9 logit units (bei 10 Trials) auf 6.25 logit units (bei 50 Trials). Die Streuung ist bei Verwendung der AFC-Paradigmen demnach anfangs größer, reduziert sich aber bei 30 und mehr Trials auf das Niveau der Ja-Nein-Abfrage.

Betrachtet man die Streuung der Fehler (in Abb. 2.11 bzw. in Tab. A.2) zeigen sich für den Bias die gleichen Abweichungen wie bei der Streuung der Schwellwerte. Ergebnisse der Messungen mit Ja-Nein-Paradigma liegen immer in der Mitte des Range, diejenigen mit AFC-Paradigma etwas ins Positive verschoben. Der Betrag der Streuung ist bei den Fehlern meist kleiner als bei den Schätzwerten selbst, da die Variation der wahren Schwellwerte nicht mehr inbegriffen ist. Die Werte sind für das Ja-Nein-Paradigma sehr gering. Der Interquartil-Abstand beträgt hier anfangs 1 logit unit und ab 30 Trials nur noch 0.5 logit units. Für die AFC-Paradigmen ergeben sich jedoch weitaus größere Unterschiede. Zum Beispiel misst man bei 10 Trials einen Abstand von 8.5 logit units zwischen den Whiskern, der sich jedoch mit steigender Anzahl von Trials stark reduziert (bis auf 2 logit units). Die Streuung der Fehler ist bei 3AFC kleiner als bei 2AFC, jedoch stets größer als bei Ja-Nein-Paradigma.

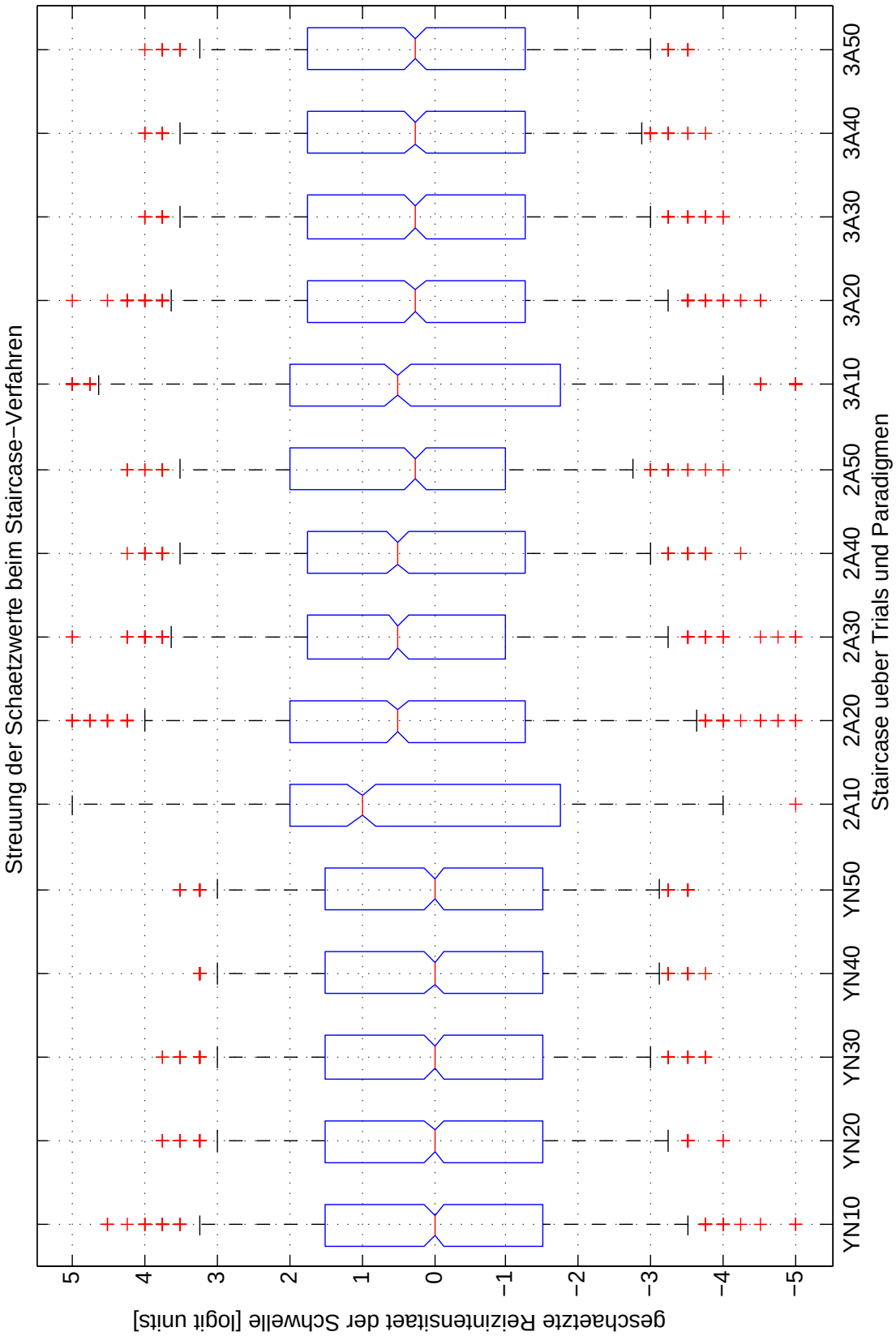

Abbildung 2.10: Verteilung der Schätzwerte beim Staircase-Verfahren

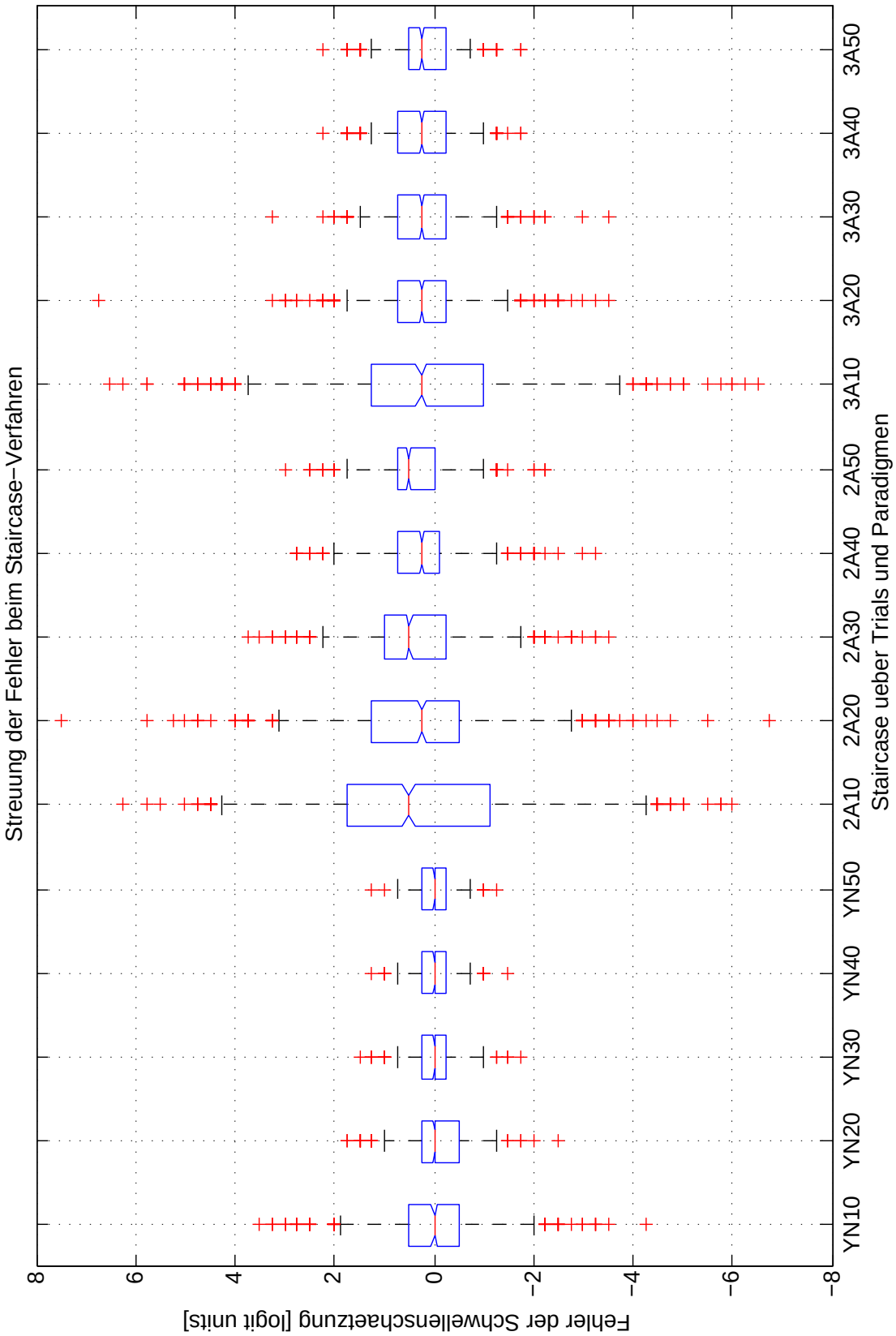

Abbildung 2.11: Verteilung der Fehler beim Staircase-Verfahren

PEST

In Abbildung 2.12 sind die Ergebnisse der Messungen mit PEST dargestellt (genauere Werte siehe Tab. A.3). Hier zeigt sich in den Medianwerten kaum Bias. Nur bei einzelnen Messungen mit AFC-Paradigma sind die Verteilungen stellenweise leicht zu negativen Werten hin verschoben. Zum Beispiel liegen die Whisker bei 2AFC und 10 Trials bei 4 und -5 logit units. Der Betrag der Streuung verringert sich nur leicht mit steigender Trialzahl. Zum Beispiel sinkt der Interquartil-Abstand bei Ja-Nein-Abfrage von 4 auf 3.5 logit units und der Bereich des 95%-Perzentil von 8 auf 6.5 logit units. Ähnliche Werte finden sich auch bei den Messungen mit AFC-Paradigma. Bei 2AFC-Paradigma ist die Streuung jedoch auch noch nach 50 Trials etwas höher als bei Ja-Nein und 3AFC.

Bei Betrachtung der Fehler (Abb. 2.13 und Tab. A.4) zeigt sich etwas deutlicher, dass die Messungen mit AFC-Paradigma von einem leichten negativen Bias geprägt sind (Median bei -0.25 logit units). Der Betrag der Streuung ist bei den Fehlern generell kleiner als bei den ermittelten Schätzwerten und zeigt für alle Paradigmen eine ähnliche Entwicklung. Die Breite der Verteilung der Fehler reduziert sich mit steigender Trialzahl sehr viel stärker als bei der Verteilung der Schätzwerte. Die Streuung ist für AFC-Paradigmen wieder etwas größer als für Ja-Nein-Abfragen.

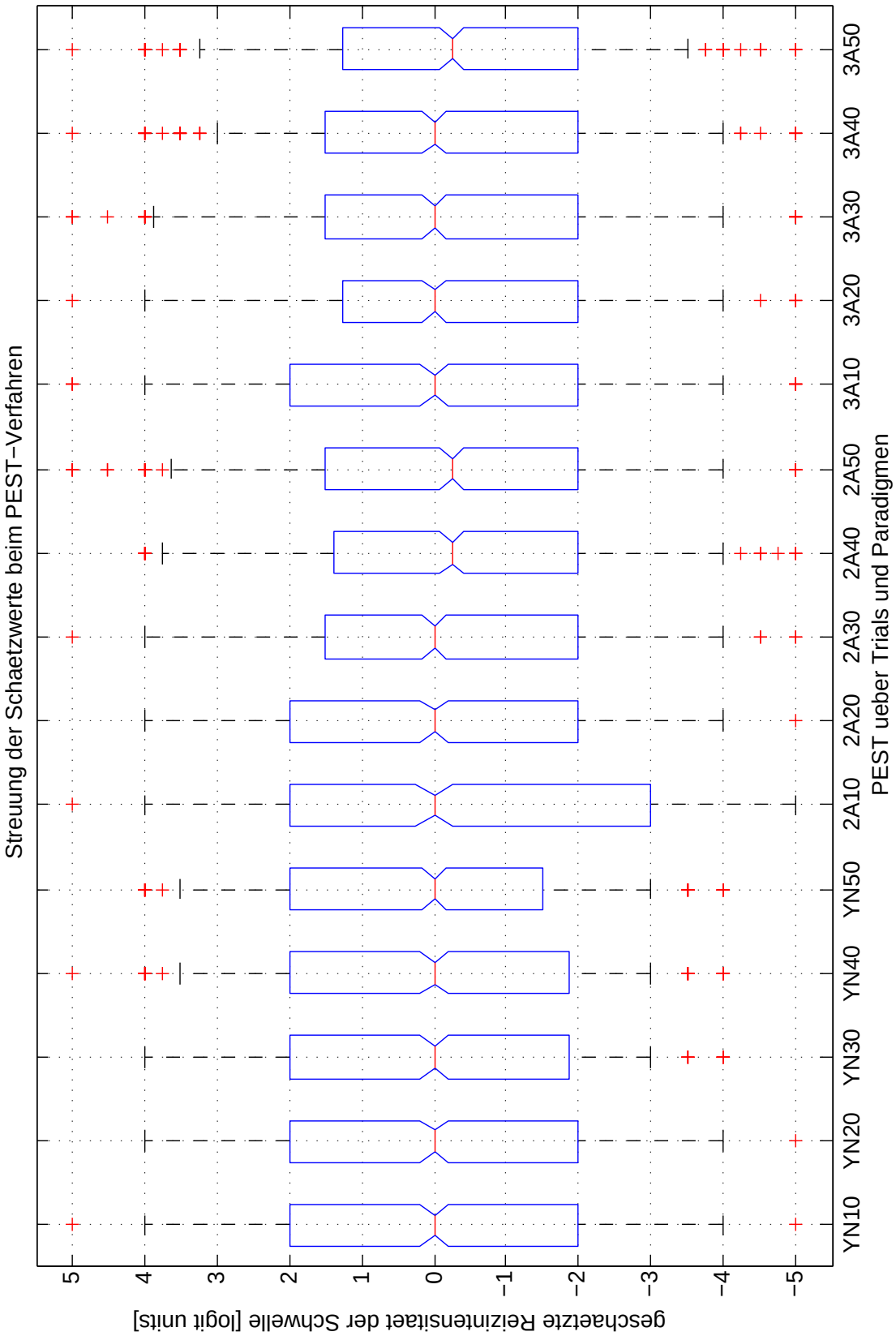

Abbildung 2.12: Verteilung der Schätzwerte beim PEST-Verfahren

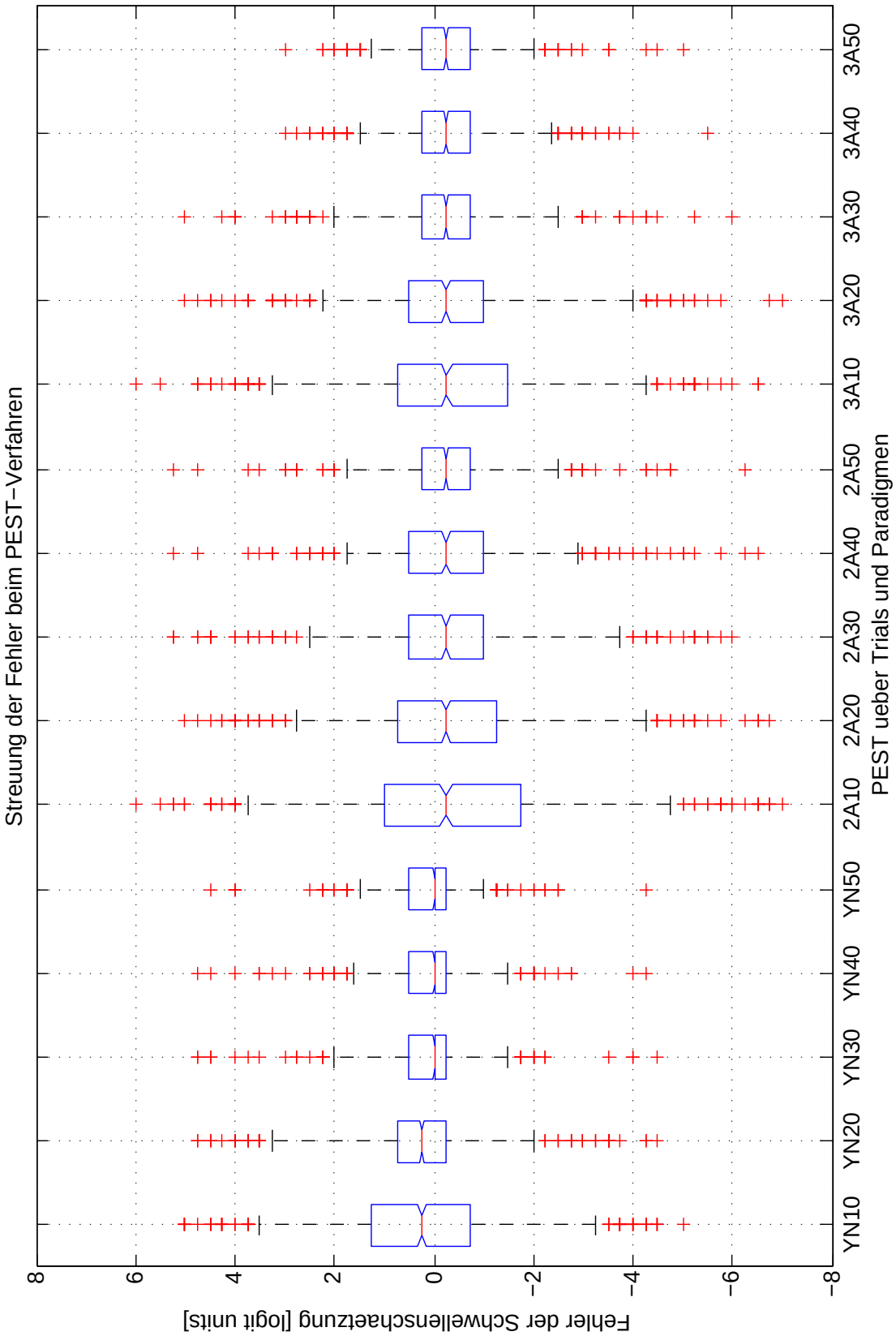

Abbildung 2.13: Verteilung der Fehler beim PEST-Verfahren

Best PEST

Abbildung 2.14 stellt die Verteilung der Schätzwerte beim Best PEST-Verfahren dar (siehe auch Tab. A.5). Bei Ja-Nein-Paradigma zeigt sich eine symmetrische Verteilung der Schätzwerte um 0 logit units. Bei Messungen mit AFC-Paradigma tritt jedoch ein erheblicher Bias auf (bis zu -1.25 logit units bei 2AFC und 10 Trials), der bei 2AFC auch noch nach 50 Trials deutlich vorhanden ist. Die Perzentile sind dabei zwar entsprechend mit verschoben, jedoch zeigt sich eine Tendenz zur linksseitigen Verteilung. Die Interquartil-Abstände verändern sich bei Best PEST mit steigender Trialzahl nur wenig (von anfangs ca. 3.5 auf 3 logit units). Betrachtet man die Länge der Whisker, zeigt sich jedoch auch hier die Tendenz zur schmaleren Verteilung (z.B. bei 3AFC von 7.75 auf 6.25 logit units).

Bei Betrachtung der Fehler-Verteilung (Abb. 2.15 und Tab. A.6) sind die gleichen Tendenzen deutlicher sichtbar. Die Messungen mit Ja-Nein-Paradigma sind frei von Bias und symmetrisch um 0 logit units verteilt. Die Ergebnisse mit AFC-Paradigma weisen teilweise starke negative Abweichungen auf, die jedoch für 3AFC mit steigender Trialzahl fast verschwinden. Die Breite der Streuung nimmt zu hohen Trialzahlen hin ab und ist für 2AFC (Interquartil-Abstand mindestens 1.5 logit units) stets größer als für 3AFC und Ja-Nein-Paradigma (Minimum jeweils bei 0.75 und 0.5 logit units).

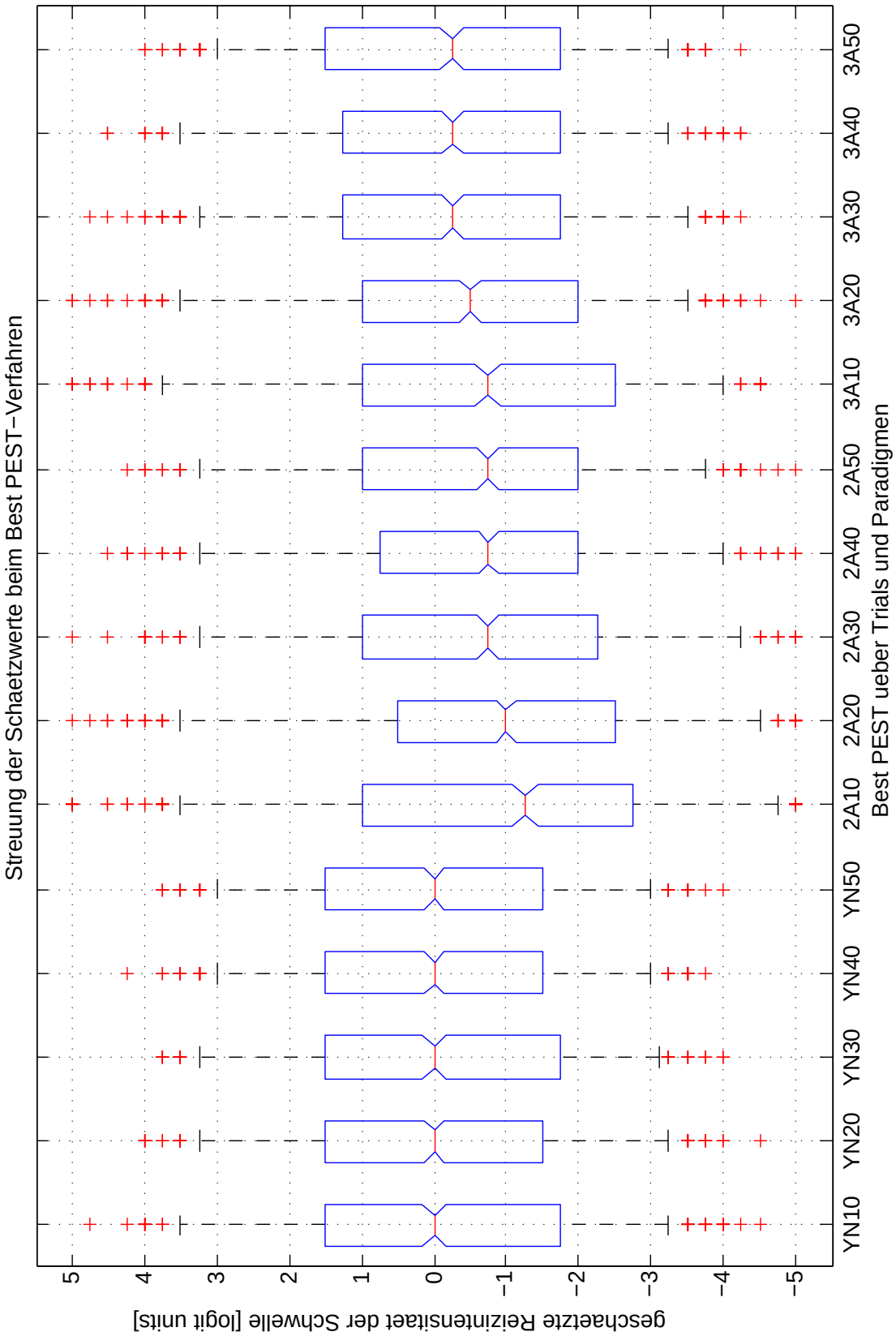

Abbildung 2.14: Verteilung der Schätzwerte beim Best PEST-Verfahren

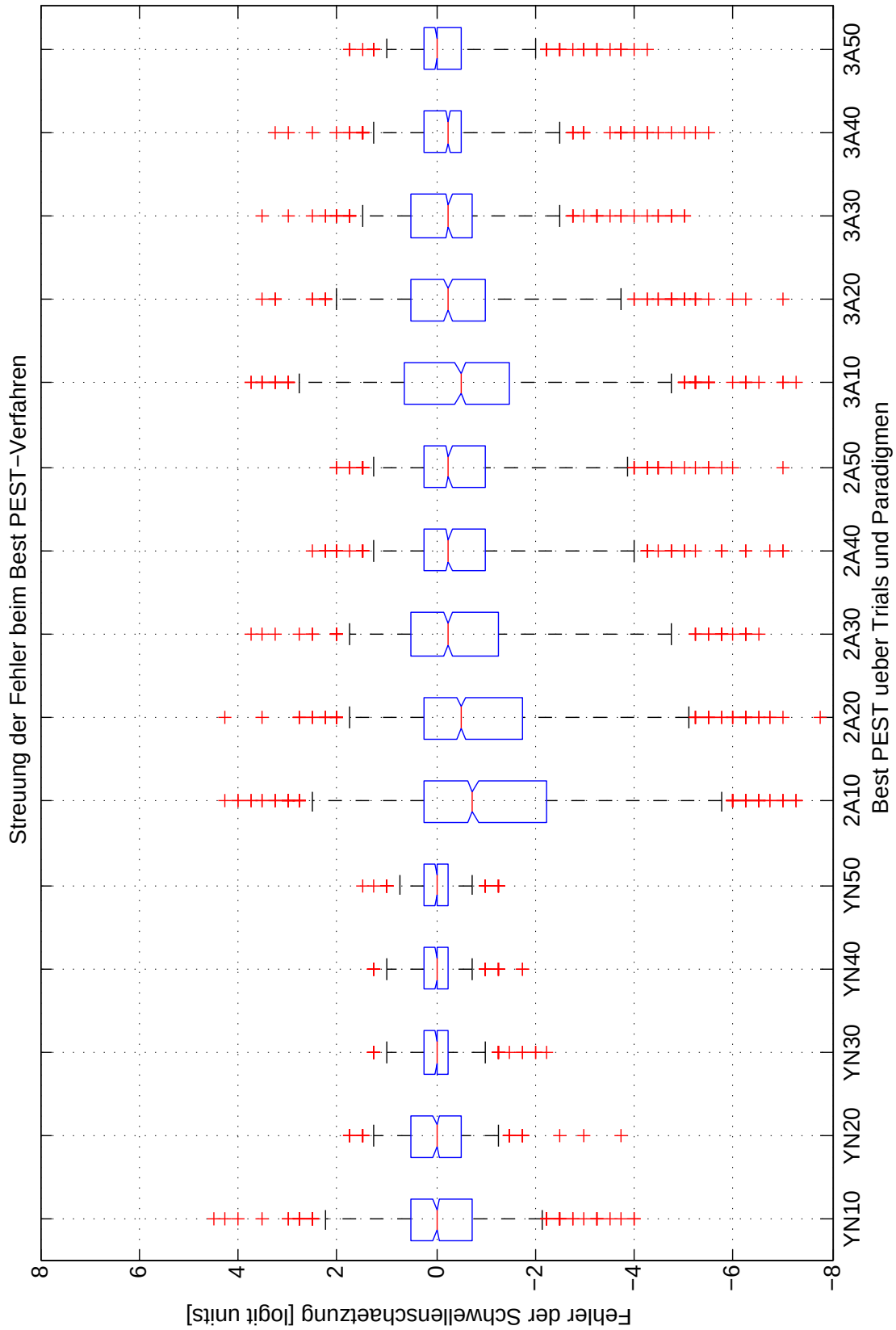

Abbildung 2.15: Verteilung der Fehler beim Best PEST-Verfahren

ZEST

In Abbildung 2.16 sind die Verteilungen der Schätzwerte beim ZEST-Verfahren dargestellt (siehe auch Tab. A.7). Bei den Ergebnissen der Evaluierung von ZEST ist kein Bias feststellbar. Die Verteilungen der Schätzwerte sowie der Fehler sind bei allen getesteten Paradigmen symmetrisch. Betrachtet man die Streuung der Schätzwerte zeigen sich mit steigender Trialzahl kaum Veränderungen. Die Interquartil-Abstände betragen meist 3 logit units, der Bereich des 95%-Perzentils umfasst anfangs 7 und bei über 30 Trials minimal 6 logit units.

Bei der Streuung der Fehler (in Abb. 2.17 und Tab. A.8) zeigt sich mit steigender Trialzahl wieder deutlich die immer schmaler werdende Verteilung. Zum Beispiel reduzieren sich die Abweichungen bei 3 AFC von anfangs 1.5 auf 0.5 logit units interquartil (bei 40 und mehr Trials). Die Fehler sind bei 2AFC etwas größer als bei 3AFC und Ja-Nein-Paradigma. Die Fehler der 3AFC-Messung sinken ab 40 Trials etwa auf das Niveau der Ergebnisse bei Ja-Nein-Abfrage.

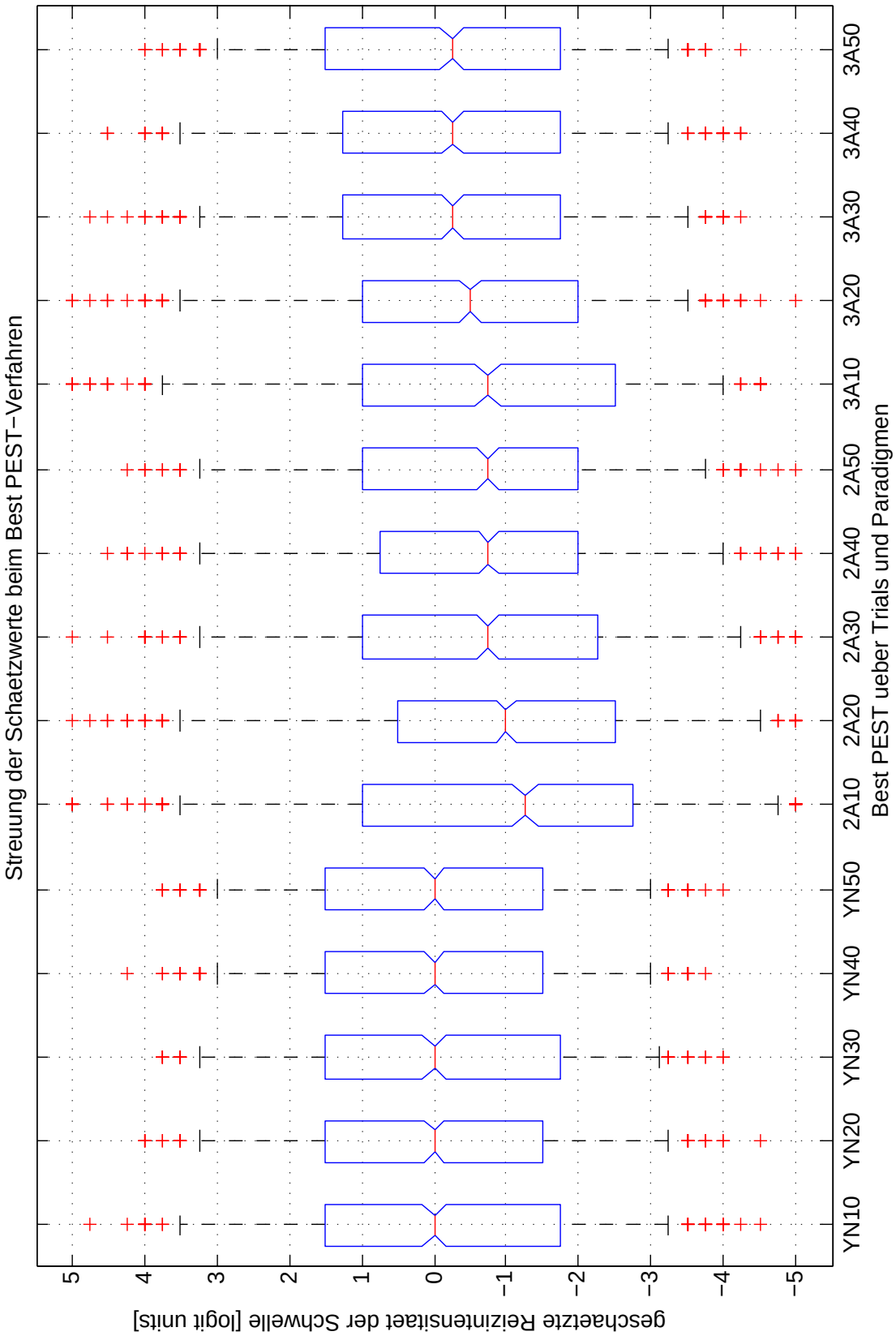

Abbildung 2.16: Verteilung der Schätzwerte beim ZEST-Verfahren

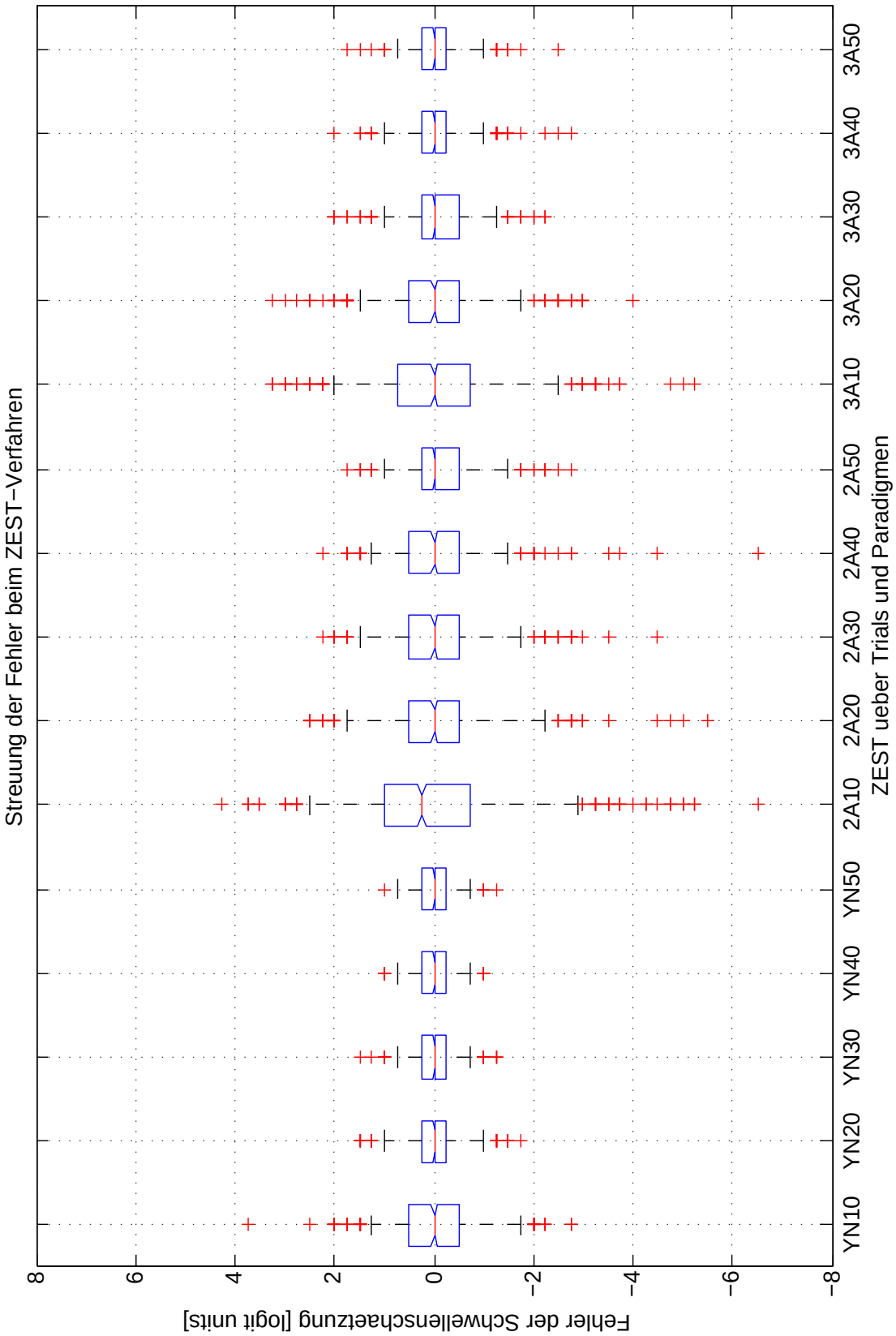

Abbildung 2.17: Verteilung der Fehler beim ZEST-Verfahren

2.3.2 Ergebnisse der Verfahren im Vergleich

Bias der Schätzwerte

Der Bias einer Messung berechnet sich als mittlerer Fehler bezogen auf die wahre Schwelle (vgl. Gl. 1.5). Dieser Wert lässt erkennen, ob eine Prozedur die Schwelle eher über- oder unterschätzt. Abbildung 2.18 zeigt den Bias der Verfahren im direkten Vergleich. Die genauen Werte können Tabelle A.9 entnommen werden. Bei den Messungen mit Ja-Nein-Abfrage ist kein Bias vorhanden. Eine Ausnahme bildet das PEST-Verfahren, dass eine konstante Überschätzung der Schwelle aufweist (Bias von 0.1 bis 0.2 logit units). Bei den AFC-Paradigmen zeigt sich ein größerer negativer Bias für das Best PEST-Verfahren (-1 bis -0.5 logit units, auch nach 50 Trials). Bei 3AFC reduziert sich diese Abweichung jedoch bis auf -0.2 logit units. Das PEST-Verfahren tendiert hier auch zur Unterschätzung der Schwelle und weist für AFC einen Bias zwischen -0.43 und -0.2 logit units auf. Die Staircase-Methode wiederum überschätzt die Schwelle um ca. 0.4 logit units bei 2AFC und 0.2 logit units bei 3AFC. Bei der ZEST-Methode kann für keines der Paradigmen Bias festgestellt werden.

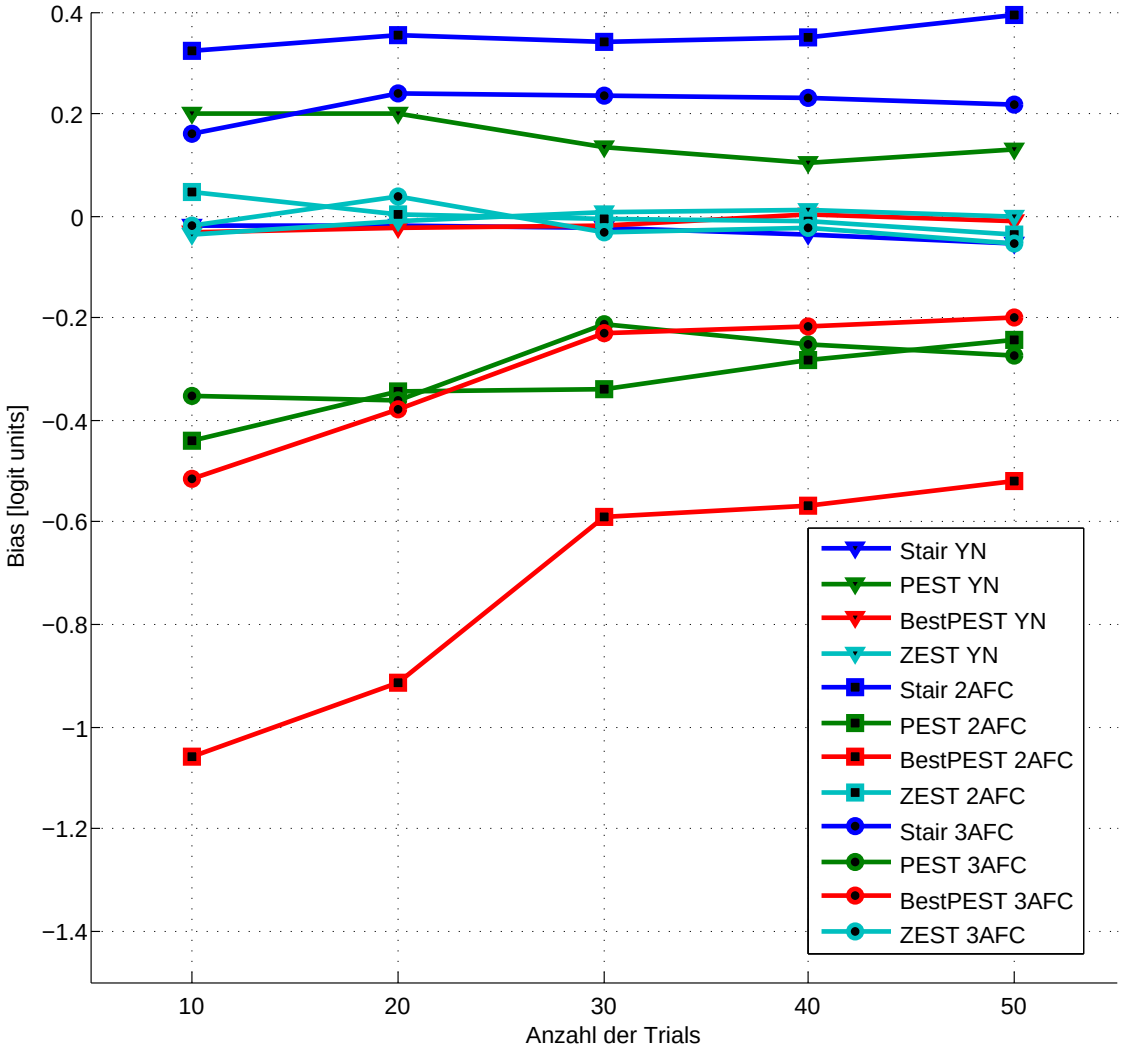

Abbildung 2.18: Bias der Verfahren im Vergleich

Varianz der Schätzwerte

Die Varianz einer Messung, dargestellt über der Anzahl von Trials, lässt erkennen wie schnell eine Prozedur konvergiert. Außerdem kann die Genauigkeit einer Messung über die Varianz abgeschätzt werden. Berechnet wird die Varianz als quadrierte Standardabweichung der Schätzwerte. In Abbildung 2.19 ist die Varianz der Schätzwerte für alle Verfahren dargestellt (siehe auch Tab. A.10). Bei den Messungen mit Ja-Nein-Paradigma zeigen sich die besten Werte der Varianz (zwischen 3.5 und 3.3 logit units^2). Staircase, Best PEST und ZEST unterscheiden sich in dieser Hinsicht nur wenig. Eine Ausnahme bildet wieder das PEST-Verfahren, das bei kleiner Trialzahl eine weitaus höhere Varianz aufweist (bei 10 Trials 7.5 logit units^2). Für AFC-Paradigmen weist ZEST sehr genaue Schätzwerte auf (anfangs 4, dann bis 3.6 logit units^2), wobei Staircase ab 30 und Best PEST ab 40 Trials ähnlich genaue Ergebnisse liefern (3.2 bis 3.8 logit units^2). Das PEST-Verfahren führt auch in Kombination mit AFC-Paradigma zur größten Varianz (von anfangs 8 bis 4.4 logit units^2).

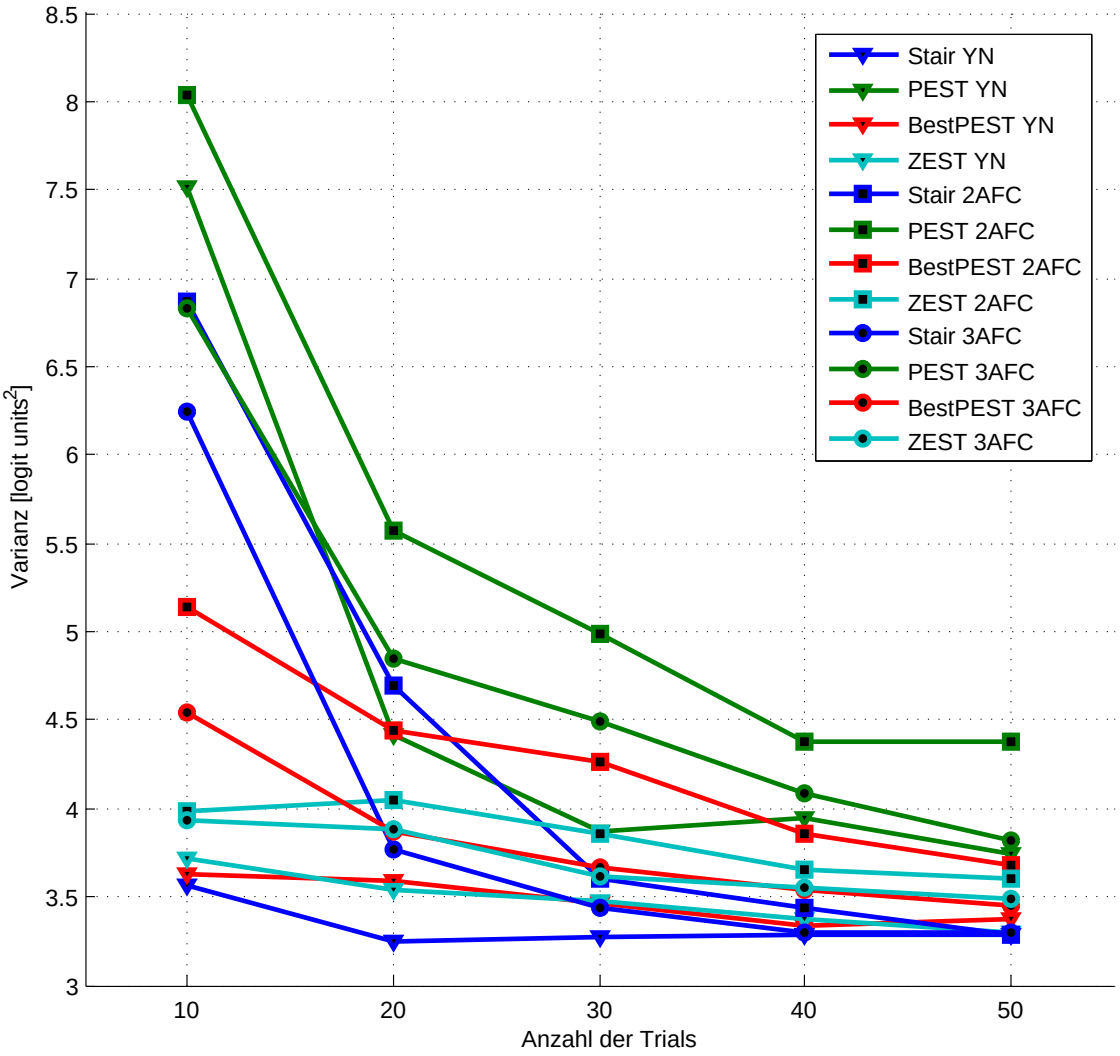

Abbildung 2.19: Varianz der Verfahren im Vergleich

Varianz der Fehler

Wie oben beschrieben werden die Fehler als Differenz zwischen wahrer und geschätzter Schwelle ermittelt. Die Varianz der wahren Schwelle wird also eliminiert. Im Folgenden wird die Varianz der Fehler für die verschiedenen Verfahren über der Anzahl der Trials dargestellt. Abbildung 2.20 stellt die Varianz der Fehler für alle Verfahren dar (siehe auch Tab. A.11). Bei jedem getesteten Paradigma weist die ZEST-Methode die geringste Varinz der Fehler auf. Sie beträgt bei Ja-Nein anfangs 0.6 und bei höherer Trialzahl nur 0.1 logit units^2. Ähnliche Werte liefern Staircase und Best PEST mit Ja-Nein-Paradigma, wobei die Abweichungen bei 10 Trials noch etwas höher liegen. PEST weist die höchste Varianz der Fehler auf. Bei 2AFC treten die größten Fehler auf. Jedoch sind die Werte bei ZEST 2AFC stets besser als die Ergebnisse von PEST Ja-Nein. Bei Staircase 2AFC reduziert sich die Varianz der Fehler nach 40 Trials auf das Niveau von ZEST (ca. 0.6 logit units^2). Best PEST und PEST verlaufen fast parallel und erreichen aber auch nach 50 Trials noch nicht das Niveau von ZEST nach 20 Trials. Bei 3AFC liegt ZEST wieder an erster Stelle. Die Staircase-Prozedur erreicht hier aber schon nach 20 Trials ein ähnliches Niveau (von 0.8 bei 20 Trials bis 0.3 logit units^2 bei 50 Trials).

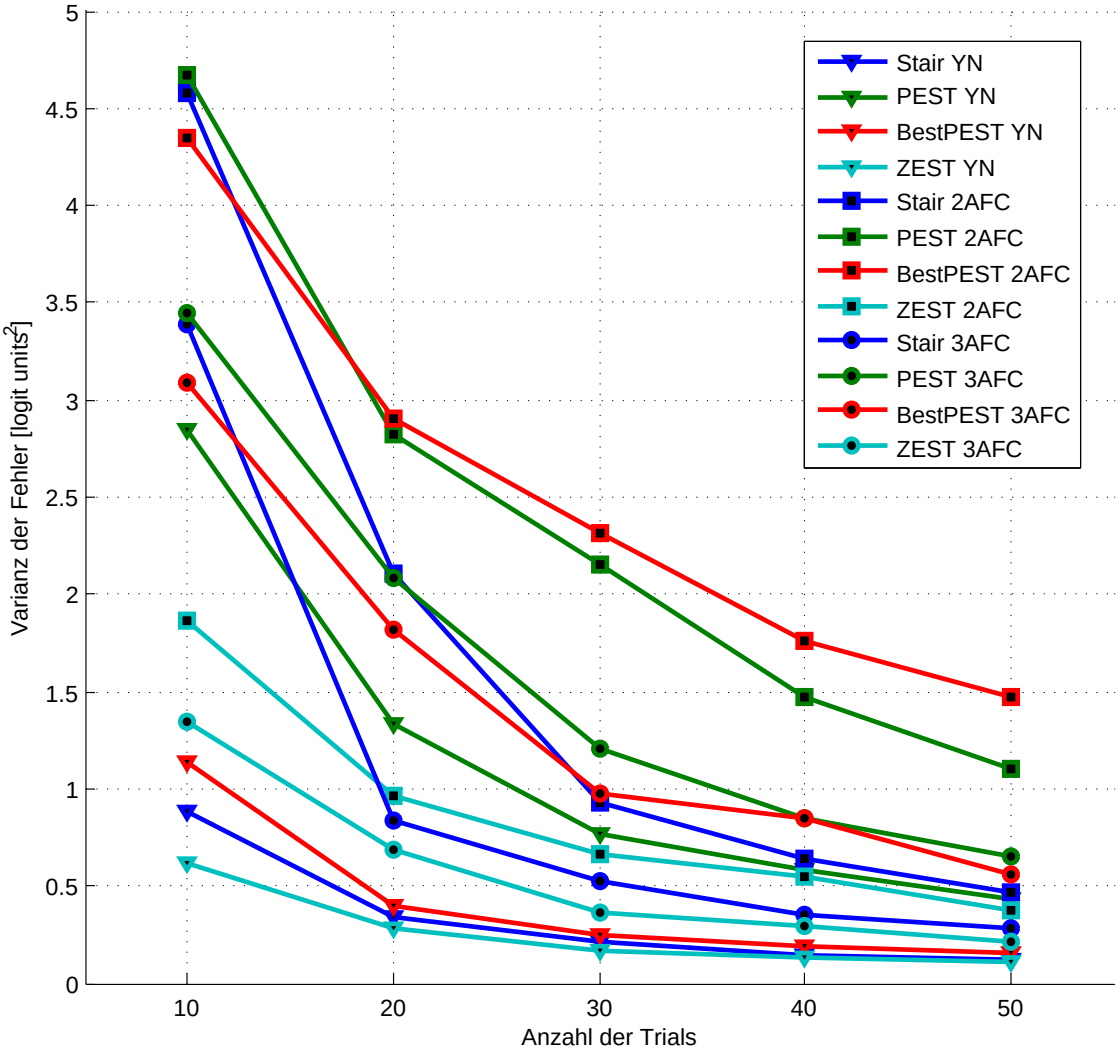

Abbildung 2.20: Varianz der Fehler im Vergleich

Sweat Factor und Effizienz

Der Sweat Faktor zeigt wie viel Aufwand bei einer Messung mit bestimmter Genauigkeit eingebracht wird und ist damit ein Indikator für die relative Effizienz der Verfahren. Berechnet wird der Sweat Factor als Produkt von Trialzahl und Varianz (vgl. Gl. 1.7). Abbildung 2.21 zeigt den Sweat Factor der Verfahren im Vergleich (siehe auch Tab. A.12). Auf den ersten Blick scheinen keine großen Unterschiede zwischen den Verfahren vorzuliegen, da die Kurven fast alle parallel verlaufen. Die Differenzen im Sweat Factor, die durch jeweils 10 weitere Trials entstehen, sind meist größer als die internen Unterschiede zwischen den verschiedenen Verfahrenstypen. Interessant sind jedoch die Schichtung der Kurven und die Punkte, an denen sich die Kurven schneiden. PEST zeigt für alle Paradigmen das schlechteste Verhältnis von Genauigkeit zu Trialzahl auf. Bei Ja-Nein Paradigma benötigt Staircase den kleinsten Aufwand, dicht gefolgt von Best PEST und ZEST. Bei den AFC-Paradigmen ist ZEST auch nach kleiner Anzahl von Trials die effizienteste Methode. Staircase und Best PEST liegen anfangs etwas höher, erreichen aber mit steigender Anzahl von Trials auch das Niveau von ZEST. Nach 50 Trial weist Staircase dann das beste Verhältnis von Genauigkeit zu Trialzahl auf.

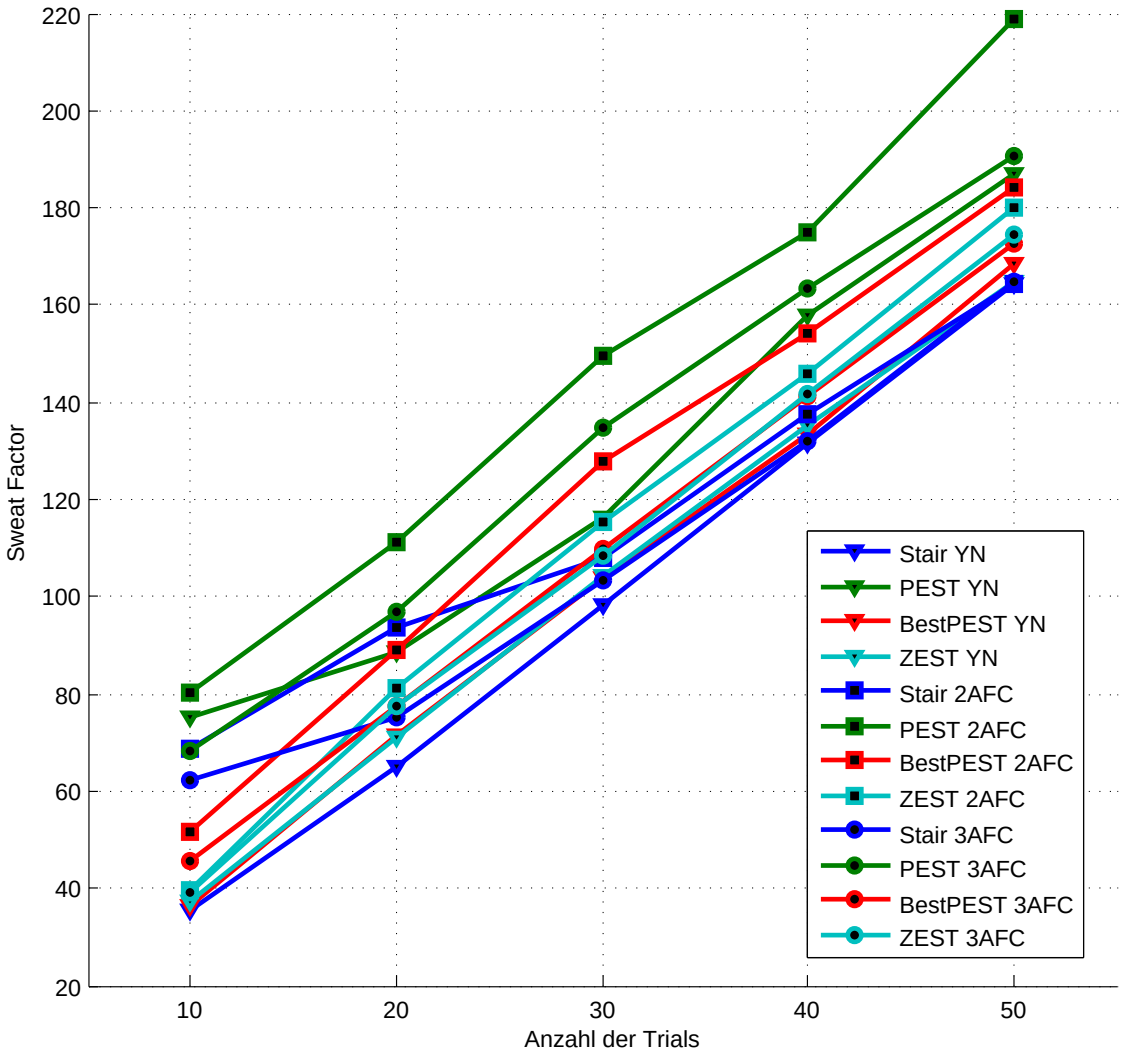

Abbildung 2.21: Sweat Factor der Verfahren im Vergleich

3 Diskussion

3.1 Auswertung der Ergebnisse

3.1.1 Staircase

Wie erwartet ist die Staircase-Methode für Ja-Nein-Paradigma frei von Bias. Bei AFC-Paradigma zeigt sich jedoch ein konstanter positiver Bias, der auf die unterschiedlichen adaptiven Regeln zurückzuführen ist, die verwendet wurden um das Konvergenzniveau in die Nähe von 50%-Korrekt zu verschieben. In den hier verwendeten Implementierungen überschätzt die Staircase-Prozedur demnach die Schwelle, wenn sie mit AFC-Paradigma kombiniert wird. Diese konstante Abweichung des Konvergenzniveaus könnte nach der Messung theoretisch herausgerechnet werden. Bei den Staircase-Varianten mit Up-Down Tranformed Response besteht weiterhin das Problem, dass das Konvergenzniveau nicht beliebig wählbar ist. Bezogen auf die Genauigkeit der Staircase-Methode wurden keine herausragenden Ergebnisse erwartet, da die anfängliche Schrittweite relativ groß gewählt wurde und eine adaptive Anpassung der Schrittweite, abgesehen von der einmaligen Halbierung, nicht möglich ist. Die tatsächlich ermittelten Werte sind aber überraschend genau, sogar so gut wie die Werte der parametrischen Verfahren. Aus Literatur und vorbereitenden Simulationen ist bekannt, dass die Genauigkeit dieser Prozedur aber stark von der Wahl der Auflösung und Schrittgröße abhängig ist. Anscheinend wurde hier eine, bezogen auf die Steigung der psychometrischen Funktion, optimale Schrittgröße gewählt. Dies bewirkt ein häufiges Pendeln im Bereich um 50%-Korrekt und führt so zu Mittelwerten nahe der wahren Schwelle. Diese Vermutung wird bekräftigt, da die Prozedur sehr schnell konvergiert und schon nach 30 Trials kaum noch Verbesserungen der Varianz auftreten. Für AFC-Paradigma und wenig Trials weisen die Ergebnisse hier etwas höhere Streuungen auf.

Dies liegt daran, dass bei dieser Triallänge nur wenig Reversals (teilweise auch gar keine Umkehrung) auftreten. Um die Mittelwertberechnung zu ermöglichen, muss in diesem Fall auch das erste Reversal in die Berechnung des resultierenden Schätzwerts mit einbezogen werden.

3.1.2 PEST

Die PEST-Methode sollte den Erwartungen nach bessere Ergebnisse liefern als die Staircase-Verfahren, jedoch größere Varianzen aufweisen als Best PEST, da die PEST-Prozedur erst spät konvergiert. Die Wahl der Anfangsschrittgröße und des Deviation Limits W hat zwar großen Einfluss auf die Effizienz, wurde aber vorher optimiert. Trotzdem liefert PEST hier überraschend schlechte Ergebnisse. Die Variante PEST Ja-Nein ist zwar genauer als Best PEST mit 3AFC, jedoch liefert PEST weitaus schlechtere Schätzwerte als die Staircase-Methode. Wahrscheinlich wurde anfangs eine zu große Schrittgröße gewählt um schon bei wenig Trials genauere Ergebnisse zu erreichen. Reversals und entsprechende Halbierungen der Schrittgröße finden erst nach 10 bis 20 Trials statt. Das langsame Konvergenzverhalten kann also hier bestätigt werden. Bei AFC zeigt sich außerdem negativer Bias, der jedoch weitgehend konstant bleibt und gegebenenfalls herausgerechnet werden könnte.

3.1.3 Best PEST

Für das Best PEST-Verfahren wurden sehr gute Schätzwerte bei Ja-Nein und etwas ungenauere Werte mit Bias bei AFC-Pradigma erwartet. Hinsichtlich des Ja-Nein-Paradigmas können die Erwartungen bestätigt werden. Für AFC zeigt das Best PEST-Verfahren jedoch sehr viel schlechtere Schätzwerte als angenommen. Auch bei dem zusätzlich evaluierten 3AFC-Paradigma ist die Genauigkeit nur so gut wie bei PEST 3AFC. Bei Ja-Nein-Abfragen ist Best PEST frei von Bias. Bei AFC tritt jedoch starker, nicht konstanter Bias auf, der auch noch nach 50 Trials erhebliche Werte aufweist. Diese Abweichungen können nur auf die Kombination von hoher Ratewahrscheinlichkeit und Maximum Likelihood-Schätzung zurückgeführt werden. Wahrscheinlich führt

wiederholtes, richtiges Raten zum „Verfangen“ der Prozedur im unterschwelligen Bereich und somit zur deutlichen Unterschätzung der Schwelle.

3.1.4 ZEST

Den Erwartungen nach sollte das ZEST-Verfahren genauer sein als Best PEST, da es nicht das Maximum sondern den Mittelwert der posterior pdf als besten Schätzwert verwendet. Deshalb sollte es auch weniger Bias aufweisen. Die Erwartungen wurden hier sogar noch übertroffen. Das ZEST-Verfahren ist für alle Paradigmen frei von Bias und liefert durchgehend sehr genaue Ergebnisse. Mit Ja-Nein liefert es die besten Schätzwerte, die Unterschiede zwischen Messungen mit verschiedenen Paradigmen sind jedoch nur minimal. Ein weiterer Vorteil ist, dass dieses Verfahren schon innerhalb der ersten 20 Trials konvergiert.

3.2 Vergleich mit früheren Simulationen

3.2.1 Vergleich mit Ergebnissen von Pentland

Abbildung 3.1 zeigt die Ergebnisse der Simulationen von Pentland (1980,[19]), die zu Staircase, PEST, improved PEST und Best PEST in Verbindung mit dem Ja-Nein-Paradigma durchgeführt wurden. Dargestellt ist die *setting accuracy* als Standardabweichung der Schätzwerte (in logit units) aufgetragen über der Anzahl von Trials. Demnach lassen sich die Ergebnisse qualitativ mit den hier ermittelten Werten für die Varianz vergleichen (siehe Abb. 2.19). Ein quantitativer Vergleich der Messwerte ist allerdings kaum möglich, da die Evaluation von Pentland nur oberflächlich dokumentiert ist. Genaue Ergebnisse sowie die für die Simulation gewählten Parameter sind nicht verfügbar. In beiden Evaluationen ist die generelle Tendenz erkennbar, dass die Streuung der Messwerte bei steigender Anzahl von Trials abnimmt. Wie bei Pentland zeigt sich auch hier bei PEST eine größere Streuung und ein langsameres Konvergenzverhalten als bei Best PEST.

Im Vergleich zur Evaluation von 1980 ist in den hier durchgeführten Messungen aber kein Nachteil für das Staircase-Verfahren nachweisbar. Es liefert mit Ja-Nein-Paradigma ähnlich gute Werte wie Best PEST. Die generelle Aussage Pentlands darüber, dass die Staircase-Methode keine besseren Schätzwerte liefern kann als PEST und Best PEST, wird infolge der hier gefunden Ergebnisse also differenzierter betrachtet.

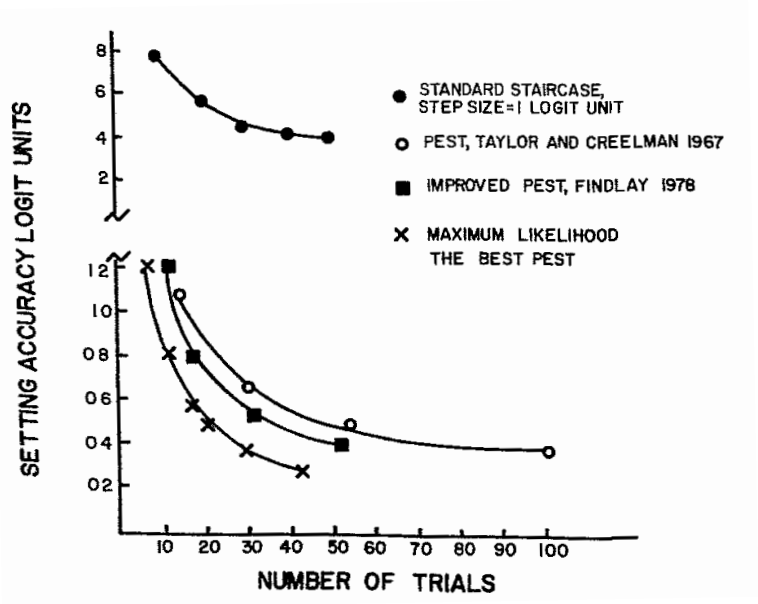

Abbildung 3.1: Ergebnisse der Simulationen von Pentland

3.2.2 Vergleich mit Ergebnissen von Madigan & Williams

In Abbildung 3.2 sind die Ergebnisse der Simulationen von Madigan & Williams dargestellt (1987,[15]). Dabei wurden PEST, Best PEST und QUEST jeweils mit Ja-Nein- und 2AFC-Paradigma kombiniert und verglichen. Im Diagramm wird die *setting accuracy* diesmal als Standardabweichung der Fehler in logit units dargestellt. Diese Werte lassen sich qualitativ mit den Fehlervarianzen aus Abb. 2.20 vergleichen. Bei Madigan & Williams weisen Best PEST und QUEST stets bessere Werte auf als PEST. Dies lässt sich hier nur für die Messung mit Ja-Nein-Paradigma bestätigen. Für AFC liegen Best PEST und PEST bei etwa gleichen Werten für die Varianz der Fehler. Bei 2AFC liefert Best PEST teilweise sogar schlechtere Schätzwerte als PEST.

Ein quantitativer Vergleich ist nur ungefähr möglich, da die Ergebnisse von Madigan & Williams nicht explizit vorliegen, sondern nur aus dem Diagramm abgelesen werden können. Außerdem wurde das PEST-Verfahren bei der früheren Evaluation mit dynamischem Abbruchkriterium durchgeführt. Die Variante Best PEST Ja-Nein weist in beiden Evaluationen ähnliche Werte auf (außer bei 10 Trials), bei Verwendung des 2AFC-Paradigmas liegen die Werte der Fehlervarianz hier jedoch höher. PEST liefert bei Madigan & Williams generell kleinere Fehlervarianzen. Dies liegt höchstwahrscheinlich daran, dass für die Simulation andere Werte für die Schrittgröße und das Deviation Limit W gewählt wurden.

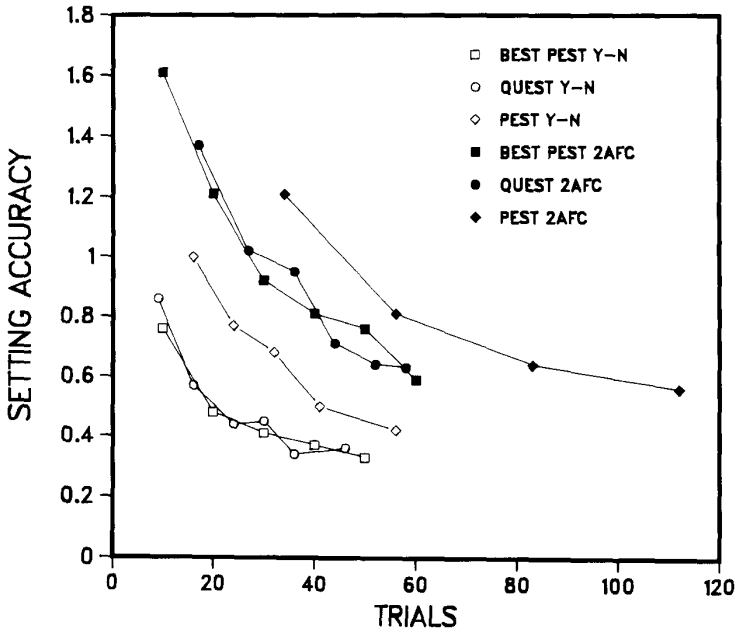

Figure 1. A comparison between maximum-likelihood psychometric procedures and Taylor and Creelman's (1967) PEST approach. Setting accuracy is the standard deviation of threshold estimation errors. Y–N = yes–no; 2AFC = two-alternative forced-choice.

Abbildung 3.2: Ergebnisse der Simulationen von Madigan & Williams

3.3 Zusammenfassung

Um biasfreie Schätzwerte zu erhalten sollte entweder das ZEST- oder Staircase-Verfahren verwendet werden. Für schnell konvergierende und genaue Messungen ist das Ja-Nein-Paradigma oder auch die ZEST-Methode zu empfehlen, die für jedes Paradigma sehr gute Schätzwerte liefert.
Die Staircase-Methode bietet den Vorteil, dass kaum Vorwissen über die zugrundeliegende psychometrische Funktion nötig ist. Allerdings ist die Genauigkeit der Schätzwerte stark von der Wahl der Auflösung und der Schrittgröße abhängig. Da Staircase einfach implementiert und durchgeführt werden kann, scheint dieses Verfahren für Vorversuche sehr empfehlenswert.
Die parametrischen Verfahren Best PEST und ZEST haben den Nachteil, dass Form und Steigung der psychometrischen Funktion schon vor der Messung ziemlich genau bekannt sein müssen, da bei der Prozedur ein psychometrisches Modell der Population zugrunde gelegt wird. Von der PEST-Methode ist generell abzuraten, da sie ungenaue Schätzwerte liefert oder sehr viele Trials benötigt um zu konvergieren. Sie könnte allerdings nützlich sein, wenn das Erreichen eines vorbestimmten Konfidenzintervalls verlangt wird.

3.4 Grenzen der Simulation und Ausblick

Die hier durchgeführten Simulationen können die bestehenden Verfahren keineswegs umfassend evaluieren. Weitere Simulationen zu anderen Verfahren und/oder Paradigmen wären nötig um die Erkenntnisse weiter zu ergänzen. Zum Beispiel könnten die hier gewählten oder ähnliche Verfahren auch mit dynamischem Abbruchkriterium implementiert und dann verglichen werden. Außerdem sollten weitergehende Simulationen zu speziellen Situationen durchgeführt werden. Dazu gehören sequentielle Abhängigkeiten wie Lerneffekte und Lapsing sowie die Wirkung von Mismatches auf die Performance der parametrischen Verfahren.

Doch auch die Methode der Simulation hat Grenzen. Simulationen liefern sehr reliable Ergebnisse, da zufällige Fehler reduziert und Einflussfaktoren kontrolliert werden

können. Bei ausreichend großem Umfang an Messungen sind die ermittelten Unterschiede stets signifikant. Um jedoch Empfehlungen für die Praxis geben zu können, müsste die Gültigkeit der Simulationsergebnisse erst noch in empirischen Tests verifiziert werden.

In realen Messungen vermischen sich die hier gefunden, verfahrenseigenen Tendenzen meist mit zufälligen Fehlern. Dies ist besonders problematisch bei Verfahren wie PEST und Best PEST, die mit Bias behaftet sind und große eigene Varianzen aufweisen. Es wäre sehr interessant zu ermitteln, ob die Unterschiede durch zufällige Fehler völlig verdeckt werden oder ob bei Verwendung der ZEST-Methode weiterhin Vorteile gegenüber anderen Verfahren nachweisbar sind.

Des Weiteren muss beachtet werden, dass Ja-Nein-Abfragen in Simulationen zwar meist sehr genaue Ergebnisse aufweisen, in realen Messungen jedoch das Kriterienproblem auftritt. Der sensorische Anteil einer Entscheidung ist dabei kaum trennbar mit dem individuell verschiedenen Entscheidungskriterium verbunden. Die Verwendung von Alternative Forced Choice-Paradigmen sollte also stets bevorzugt werden.

Die Staircase-Methode liefert in den hier durchgeführten Simulationen sehr gute Ergebnisse. Der Verlauf dieser sehr einfach konzipierten Messmethode ist von realen Versuchspersonen jedoch leicht durchschaubar und demzufolge auch manipulierbar. Verfahren wie ZEST sind in dieser Hinsicht vorzuziehen, da die adaptive Anpassung der Schrittgröße und die Richtungsänderung der Reizstärke nicht unmittelbar von der letzten Antwort abhängig sind. Es besteht jedoch weiterhin das Problem der Berücksichtigung sequentieller Abhängigkeiten. Lerneffekte, Ermüdung und Konzentration sind genauso schwer simulierbar wie das Rateverhalten von Versuchspersonen, da diese Phänomene keinen Regeln folgen, aber auch nicht völlig zufällig auftreten. Der Faktor „Versuchsperson“ darf also nicht unterschätzt werden.

Literaturverzeichnis

[1] Amitay, S., Irwin, A., Hawkey, D., Cowan, J. & Moore, D. (2006). A comparison of adaptive procedures for rapid and reliable threshold assessment and training in naive listeners. JASA, 119, 1616-1625

[2] Dixon, W.J. & Mood, A.M. (1948). A method for obtaining and analyzing sensitivity data. Journal of the American Statistical Association, 43, 109-126

[3] Emerson, P.L. (1986a). Observations on maximum-likelihood and Bayesian methods of forced-choice sequential threshold estimation. Perception & Psychophysics, 39(2), 151-153

[4] Fechner, G.T. (1860). Elemente der Psychophysik (Vol.1). Leipzig: Breitkopf und Härtel

[5] Findlay, J.M. (1978). Estimates on probability functions: A more virulent PEST. Perception & Psychophysics, 23, 181-185

[6] Gelfand, S.A. (2004). Hearing. An Introduction to psychological and physiological acoustics. 4. überarb. u. erw. Auflage, New York: Dekker

[7] Green, D.M. & Swets, J.A. (1966). Signal Detection Theory and Psychophysics. New York: Wiley

[8] King-Smith, P.E., Grigsby, S.S., Vingrys, A.J., Benes, S.C. & Supowit, A. (1994). Efficient and Unbiased Modifications of the QUEST Threshold Method: Theory, Simulations, Experimental Evaluation and Practical Implementation. Vision Research, 34, 885-912

[9] Kollmeier, B., Gilkey, R.H. & Sieben, U.K. (1988). Adaptive staircase techniques in psychoacoustics: A comparison of human data and a mathematical model. JASA, 83, 1852-1862

[10] Laming, D. & Marsh, D. (1988). Some performance tests of QUEST on measurements of vibrotactile thresholds. Perception & Psychophysics, 44, 99-107

[11] Levitt, H. (1971). Transformed up-down methods in psychoacoustics. JASA, 49, 467-477

[12] Lieberman, H.R. & Pentland, A.P. (1982). Computer Tecnology. Microcomputer-based estimation of psychophysical thresholds: the Best PEST. Behavior Research Methods & Instrumentation, 14, 21-25

[13] Luce, R.D. (1959). Individual choice behavior. New York: Wiley

[14] Luce, R.D. (1963). Detection and recognition. In Luce, R.D., Bush, R.R. & Galanter, E., eds, Handbook of mathematical psychology, volume 1, S. 103-189, New York: Wiley

[15] Madigan, R. & Williams, D. (1987). Maximum-likelihood psychometric procedures in two-alternative forced-choice: Evaluation and recommendations. Perception & Psychophysics, 42, 240-249

[16] Maempel, H.-J. (2007). Perzeptive Messung und Evaluation. Skript zum Labor Kommunikationstechnik I/II, WS 2006/2007

[17] Marvit, P., Florentine, M. and Buus, S. (2003). A comparison of psychophysical procedures for level-discrimination thresholds. JASA, 113, 3348-3361

[18] Pelli, D.G. (1987). The ideal psychometric procedure. Investigative Ophthalmology and Visual Science (Suppl.), 28, 336

[19] Pentland, A. (1980). Maximum Likelihood Estimation: the best PEST. Perception & Psychophysics, 28, 377-379

[20] Remus, J.J. & Collins, L.M. (2008). Comparison of adaptive psychometric procedures motivated by the Theory of Optimal Experiments: Simulated and experimental results. JASA. 123, 315-326

[21] Robbins, H. & Monro, S. (1951). A stochastic approximation method. Annals of Mathematical Statistics, 22, 400-407

[22] Schlauch, R.S. & Rose, R.M. (1990) Two-, three- and four-interval forced-choice staircase procedures: Estimator bias and efficiency. JASA, 88, 732-740

[23] Shelton, B.R., Picardi, M.C. & Green, D.M. (1982). Comparison of three adaptive psychophysical procedures. JASA, 71, 1527-1533

[24] Taylor, M.M. (1971). On the efficiency of psychophysical measurement. JASA, 49, 505-508

[25] Taylor, M.M. & Creelman, C.D. (1967). PEST: Efficient Estimates on probability functions. JASA, 41, 782-787

[26] Treutwein, B. (1995). Adaptive Psychophysical Procedures. Vision Research, 35, 2503-2522

[27] Watson, A.B. & Pelli, D.P. (1983). QUEST: A Bayesian adaptive psychometric method. Perception & Psychophysics, 33, 113-120

A Tabellenanhang

A.1 Daten zu Boxplots der Schätzwert- und Fehlerstreuung

In den folgenden Tabellen sind die Perzentile der Schätzwerte und die Perzentile der Fehler (in [logit units]) aufgeführt. Außerdem wurden unteres und oberes Quartil zum Interquartil-Abstand sowie 2,5% und 97,5%-Wert zum 95%-Intervall zusammengefasst.

A.1.1 Ergebnisse Staircase

Verteilungen der Schätzwerte

Bedingung	2,5%	25%	Median	75%	97,5%	Interquartil	95%-Intervall
YN 10	-3,5	-1,5	0	1,5	3,25	3	6,75
YN 20	-3,25	-1,5	0	1,5	3	3	6,25
YN 30	-3	-1,5	0	1,5	3	3	6
YN 40	-3,125	-1,5	0	1,5	3	3	6,125
YN 50	-3,125	-1,5	0	1,5	3	3	6,125
2AFC 10	-4	-1,75	1	2	5	3,75	9
2AFC 20	-3,625	-1,25	0,5	2	4	3,75	7,375
2AFC 30	-3,25	-1	0,5	1,75	3,625	2,75	6,875
2AFC 40	-3	-1,25	0,5	1,75	3,5	3	6,5
2AFC 50	-2	-1	0,25	2	3,5	3	5,5
3AFC 10	-4	-1,75	0,5	2	4,625	3,75	8,625
3AFC 20	-3,25	-1,25	0,25	1,75	3,625	3	6,875
3AFC 30	-3	-1,25	0,25	1,75	3,5	3	6,5
3AFC 40	-2,875	-1,25	0,25	1,75	3,5	3	6,375
3AFC 50	-3	-1,25	0,25	1,75	3,25	3	6,25

Tabelle A.1: Perzentile der Schätzwerte bei Staircase

Verteilungen der Fehler

Bedingung	2,5%	25%	Median	75%	97,5%	Interquartil	95%-Intervall
YN 10	-2	-0,5	0	0,5	1,875	1	3,875
YN 20	-1,25	-0,5	0	0,25	1	0,75	2,25
YN 30	-1	-0,25	0	0,25	0,75	0,5	1,75
YN 40	-0,75	-0,25	0	0,25	0,75	0,5	1,5
YN 50	-0,75	-0,25	0	0,25	0,75	0,5	1,5
2AFC 10	-4,25	-1,125	0,5	1,75	4,25	2,875	8,5
2AFC 20	-2,75	-0,5	0,25	1,25	3,125	1,75	5,875
2AFC 30	-1,75	-0,25	0,5	1	2,25	1,25	4
2AFC 40	-1,25	-0,125	0,25	0,75	2	0,875	3,25
2AFC 50	-1	0	0,5	0,75	1,75	0,75	2,75
3AFC 10	-3,75	-1	0,25	1,25	3,75	2,25	7,5
3AFC 20	-1,5	-0,25	0,25	0,75	1,75	1	3,25
3AFC 30	-1,25	-0,25	0,25	0,75	1,5	1	2,75
3AFC 40	-1	-0,25	0,25	0,75	1,25	1	2,25
3AFC 50	-0,75	-0,25	0,25	0,5	1,25	0,75	2

Tabelle A.2: Perzentile der Fehler bei Staircase

A.1.2 Ergebnisse PEST

Verteilungen der Schätzwerte

Bedingung	2,5%	25%	Median	75%	97,5%	Interquartil	95%-Intervall
YN 10	-4	-2	0	2	4	4	8
YN 20	-4	-2	0	2	4	4	8
YN 30	-3	-1,875	0	2	4	3,875	7
YN 40	-3	-1,875	0	2	3,5	3,875	6,5
YN 50	-3	-1,5	0	2	3,5	3,5	6,5
2AFC 10	-5	-3	0	2	4	5	9
2AFC 20	-4	-2	0	2	4	4	8
2AFC 30	-4	-2	0	1,5	4	3,5	8
2AFC 40	-4	-2	-0,25	1,375	3,75	3,375	7,5
2AFC 50	-4	-2	-0,25	1,5	3,625	3,5	7,625
3AFC 10	-4	-2	0	2	4	4	8
3AFC 20	-4	-2	0	1,25	4	3,25	8
3AFC 30	-4	-2	0	1,5	3,875	3,5	7,875
3AFC 40	-4	-2	0	1,5	3	3,5	7
3AFC 50	-3,5	-2	-0,25	1,25	3,25	3,25	6,75

Tabelle A.3: Perzentile der Schätzwerte bei PEST

Verteilungen der Fehler

Bedingung	2,5%	25%	Median	75%	97,5%	Interquartil	95%-Intervall
YN 10	-3,25	-0,75	0,25	1,25	3,5	2	6,75
YN 20	-2	-0,25	0,25	0,75	3,25	1	5,25
YN 30	-1,5	-0,25	0	0,5	2	0,75	3,5
YN 40	-1,5	-0,25	0	0,5	1,625	0,75	3,125
YN 50	-1	-0,25	0	0,5	1,5	0,75	2,5
2AFC 10	-4,75	-1,75	-0,25	1	3,75	2,75	8,5
2AFC 20	-4,25	-1,25	-0,25	0,75	2,75	2	7
2AFC 30	-3,75	-1	-0,25	0,5	2,5	1,5	6,25
2AFC 40	-2,875	-1	-0,25	0,5	1,75	1,5	4,625
2AFC 50	-2,5	-0,5	-0,25	0,25	1,75	0,75	4,25
3AFC 10	-4,25	-1,5	-0,25	0,75	3,25	2,25	7,5
3AFC 20	-4	-1	-0,25	0,5	2,25	1,5	6,25
3AFC 30	-2,5	-0,75	-0,25	0,25	2	1	4,5
3AFC 40	-2,375	-0,75	-0,25	0,25	1,5	1	3,875
3AFC 50	-2	-0,75	-0,25	0,25	1,25	1	3,25

Tabelle A.4: Perzentile der Fehler bei PEST

A.1.3 Ergebnisse Best PEST

Verteilungen der Schätzwerte

Bedingung	2,5%	25%	Median	75%	97,5%	Interquartil	95%-Intervall
YN 10	-3,25	-1,75	0	1,5	3,5	3,25	6,75
YN 20	-3,25	-1,5	0	1,5	3,25	3	6,5
YN 30	-3,125	-1,75	0	1,5	3,25	3,25	6,375
YN 40	-3	-1,5	0	1,5	3	3	6
YN 50	-3	-1,5	0	1,5	3	3	6
2AFC 10	-4,75	-2,75	-1,25	1	3,5	3,75	8,25
2AFC 20	-4,5	-2,5	-1	0,5	3,5	3	8
2AFC 30	-4,25	-2,25	-0,75	1	3,25	3,25	7,5
2AFC 40	-4	-2	-0,75	0,75	3,25	2,75	7,25
2AFC 50	-3,75	-2	-0,75	1	3,25	3	7
3AFC 10	-4	-2,5	-0,75	1	3,75	3,5	7,75
3AFC 20	-3,5	-2	-0,5	1	3,5	3	7
3AFC 30	-3,5	-1,75	-0,25	1,25	3,25	3	6,75
3AFC 40	-3,25	-1,75	-0,25	1,25	3,5	3	6,75
3AFC 50	-3,25	-1,75	-0,25	1,5	3	3,25	6,25

Tabelle A.5: Perzentile der Schätzwerte bei Best PEST

Verteilungen der Fehler

Bedingung	2,5%	25%	Median	75%	97,5%	Interquartil	95%-Intervall
YN 10	-2,125	-0,75	0	0,5	2,25	1,25	4,375
YN 20	-1,25	-0,5	0	0,5	1,25	1	2,5
YN 30	-1	-0,25	0	0,25	1	0,5	2
YN 40	-0,75	-0,25	0	0,25	1	0,5	1,75
YN 50	-0,75	-0,25	0	0,25	0,75	0,5	1,5
2AFC 10	-5,75	-2,25	-0,75	0,25	2,5	2,5	8,25
2AFC 20	-5,125	-1,75	-0,5	0,25	1,75	2	6,875
2AFC 30	-4,75	-1,25	-0,25	0,5	1,75	1,75	6,5
2AFC 40	-4	-1	-0,25	0,25	1,25	1,25	5,25
2AFC 50	-3,875	-1	-0,25	0,25	1,25	1,25	5,125
3AFC 10	-4,75	-1,5	-0,5	0,625	2,75	2,125	7,5
3AFC 20	-3,75	-1	-0,25	0,5	2	1,5	5,75
3AFC 30	-2,5	-0,75	-0,25	0,5	1,5	1,25	4
3AFC 40	-2,5	-0,5	-0,25	0,25	1,25	0,75	3,75
3AFC 50	-2	-0,5	0	0,25	1	0,75	3

Tabelle A.6: Perzentile der Fehler bei Best PEST

A.1.4 Ergebnisse ZEST

Verteilungen der Schätzwerte

Bedingung	2,5%	25%	Median	75%	97,5%	Interquartil	95%-Intervall
YN 10	-3,5	-1,5	0	1,5	3,5	3	7
YN 20	-3,25	-1,5	0	1,5	3,25	3	6,5
YN 30	-3	-1,5	0	1,5	3	3	6
YN 40	-3	-1,5	0	1,5	3	3	6
YN 50	-3	-1,5	0	1,5	3	3	6
2AFC 10	-3,5	-1,75	0,25	1,75	3,5	3,5	7
2AFC 20	-3,5	-1,5	0	1,5	3,5	3	7
2AFC 30	-3,5	-1,75	0	1,5	3,5	3,25	7
2AFC 40	-3,375	-1,5	-0,25	1,5	3,25	3	6,875
2AFC 50	-3,25	-1,5	0	1,5	3,25	3	6,5
3AFC 10	-3,5	-1,75	0	1,5	3,5	3,25	7
3AFC 20	-3,5	-1,5	0	1,625	3,5	3,125	7
3AFC 30	-3,5	-1,5	0	1,5	3,25	3	6,75
3AFC 40	-3,25	-1,5	0	1,5	3,125	3	6,375
3AFC 50	-3,25	-1,75	0	1,5	3,125	3,25	6,375

Tabelle A.7: Perzentile der Schätzwerte bei ZEST

Verteilungen der Fehler

Bedingung	2,5%	25%	Median	75%	97,5%	Interquartil	95%-Intervall
YN 10	-1,75	-0,5	0	0,5	1,25	1	3
YN 20	-1	-0,25	0	0,25	1	0,5	2
YN 30	-0,75	-0,25	0	0,25	0,75	0,5	1,5
YN 40	-0,75	-0,25	0	0,25	0,75	0,5	1,5
YN 50	-0,75	-0,25	0	0,25	0,75	0,5	1,5
2AFC 10	-2,875	-0,75	0,25	1	2,5	1,75	5,375
2AFC 20	-2,25	-0,5	0	0,5	1,75	1	4
2AFC 30	-1,75	-0,5	0	0,5	1,5	1	3,25
2AFC 40	-1,5	-0,5	0	0,5	1,25	1	2,75
2AFC 50	-1,5	-0,5	0	0,25	1	0,75	2,5
3AFC 10	-2,5	-0,75	0	0,75	2	1,5	4,5
3AFC 20	-1,75	-0,5	0	0,5	1,5	1	3,25
3AFC 30	-1,25	-0,5	0	0,25	1	0,75	2,25
3AFC 40	-1	-0,25	0	0,25	1	0,5	2
3AFC 50	-1	-0,25	0	0,25	0,75	0,5	1,75

Tabelle A.8: Perzentile der Fehler bei ZEST

A.2 Daten zu den Diagrammen der Bewertungsgrößen

In den folgenden Tabellen sind die Werte der ermittelten Bewertungsgrößen Bias, Varianz, Fehlervarianz und SweatFactor dargestellt.

A.2.1 Bias

Verfahren	10 Trials	20 Trials	30 Trials	40 Trials	50 Trials
YN Staircase	-0,0175	-0,0195	-0,0213	-0,0380	-0,0545
YN PEST	0,2010	0,2018	0,1328	0,1040	0,1298
YN Best PEST	-0,0325	-0,0250	-0,0203	0,0032	-0,0083
YN ZEST	-0,0365	-0,0088	0,0080	0,0095	-0,0005
2AFC Staircase	0,3220	0,3523	0,3390	0,3490	0,3917
2AFC PEST	-0,4390	-0,3443	-0,3397	-0,2802	-0,2430
2AFC Best PEST	-1,0588	-0,9133	-0,5877	-0,5667	-0,5188
2AFC ZEST	0,0450	0,0020	-0,0053	-0,0107	-0,0360
3AFC Staircase	0,1615	0,2415	0,2352	0,2293	0,2182
3AFC PEST	-0,3545	-0,3600	-0,2102	-0,2520	-0,2730
3AFC Best PEST	-0,5155	-0,3795	-0,2285	-0,2160	-0,2003
3AFC ZEST	-0,0205	0,0390	-0,0323	-0,0215	-0,0548

Tabelle A.9: Bias der Verfahren [logit units]

A.2.2 Varianz

Verfahren	10 Trials	20 Trials	30 Trials	40 Trials	50 Trials
YN Staircase	3,5650	3,2480	3,2793	3,2880	3,2849
YN PEST	7,5121	4,4199	3,8706	3,9469	3,7372
YN Best PEST	3,6237	3,5862	3,4574	3,3346	3,3731
YN ZEST	3,7213	3,5443	3,4702	3,3805	3,2982
2AFC Staircase	6,8714	4,6938	3,6057	3,4426	3,2826
2AFC PEST	8,0403	5,5704	4,9887	4,3713	4,3760
2AFC Best PEST	5,1449	4,4474	4,2671	3,8600	3,6814
2AFC ZEST	3,9841	4,0533	3,8521	3,6487	3,6014
3AFC Staircase	6,2472	3,7707	3,4409	3,3041	3,2928
3AFC PEST	6,8254	4,8427	4,4976	4,0852	3,8153
3AFC Best PEST	4.5459	3,8688	3,6622	3,5361	3,4569
3AFC ZEST	3,9301	3,8769	3,6183	3,5462	3,4888

Tabelle A.10: Varianz der Verfahren [logit units^2]

A.2.3 Varianz der Fehler

Verfahren	10 Trials	20 Trials	30 Trials	40 Trials	50 Trials
YN Staircase	0,8853	0,3382	0,2119	0,1470	0,1238
YN PEST	2,8504	1,3284	0,7630	0,5808	0,4380
YN Best PEST	1,1346	0,4037	0,2487	0,1915	0,1534
YN ZEST	0,6145	0,2876	0,1692	0,1295	0,1114
2AFC Staircase	4,5825	2,1075	0,9265	0,6411	0,4686
2AFC PEST	4,6699	2,8204	2,1585	1,4690	1,1051
2AFC Best PEST	4,3438	2,8988	2,3097	1,7640	1,4734
2AFC ZEST	1,8621	0,9697	0,6653	0,5452	0,3771
3AFC Staircase	3,3921	0,8411	0,5284	0,3487	0,2817
3AFC PEST	3,4414	2,0827	1,2009	0,8472	0,6512
3AFC Best PEST	3,0921	1,8153	0,9749	0,8452	0,5568
3AFC ZEST	1,3497	0,6875	0,3665	0,2920	0,2079

Tabelle A.11: Varianz der Fehler der Verfahren [logit units^2]

A.2.4 Sweat Factor

Verfahren	10 Trials	20 Trials	30 Trials	40 Trials	50 Trials
YN Staircase	35,6501	64,9599	98,3780	131,5188	164,2470
YN PEST	75,1211	88,3981	116,1168	157,8752	186,8607
YN Best PEST	36,2369	71,7242	103,7208	133,3855	168,6559
YN ZEST	37,2126	70,8856	104,1047	135,2216	164,9086
2AFC Staircase	68,7144	93,8760	108,1705	137,7057	164,1314
2AFC PEST	80,4032	111,4076	149,6613	174,8758	218,7976
2AFC Best PEST	51,4488	88,9472	128,0139	154,3987	184,0696
2AFC ZEST	39,8408	81,0660	115,5641	145,9488	180,0715
3AFC Staircase	62,4716	75,4190	103,2273	132,1624	164,6424
3AFC PEST	68,2541	96,8549	134,9282	163,4082	190,7643
3AFC Best PEST	45,4593	77,3770	109,8660	141,4452	172,8460
3AFC ZEST	39,3013	77,5371	108,5480	141,8484	174,4402

Tabelle A.12: Sweat Factor der Verfahren [logit units^2]

B Matlab-Files

Im folgenden Abschnitt werden die Quellcodes zur Implementierung der Verfahren und zur Auswertung der gewonnenen Daten aufgeführt. Zur Staircase-Methode ist hier nur die 1Down-1Up-Variante dargestellt. Die Quellcodes zu 2Down-1Up und 3Down-1Up finden sich auf der beiliegenden CD im Ordner *Implementierung*.

B.1 Implementierung von Staircase

```
% ========================================================================= %
% Simple Staircase Simulation nach Levitt
% ========================================================================= %
clc; clear all; close all;

% Initialization
%----------------

% Staircase = Methode 1
Methode = 1;
% Yes-No = 1, 2AFC = 2, 3AFC = 3
Para = 1;

% Anzahl der Durchläufe, Range und Trialzahl
NumRuns = 1000;
maxRange = 5;
NumTrials = 10;

pg = 0.03;    % guessing rate
pl = 0.02;    % lapsing rate
InitialStepSize = 2;
% start calculation of final threshold at first or second turnaroundpoint?
StartMean = 2;
RealSlope = 2*maxRange/10;

% Definieren des leeren Daten-Arrays
Data = struct(  'Method', Methode, ...
                'Paradigm', Para,  ...
                'Trials', NumTrials,  ...
                'Stimuli', zeros(NumTrials,NumRuns), ...
                'Responses', zeros(NumTrials,NumRuns), ...
                'RealThresholds', zeros(1,NumRuns), ...
                'EstimateThresholds', zeros(1,NumRuns), ...
                'Sonstiges', [maxRange, pg, pl, RealSlope, InitialStepSize, ...
                 StartMean, zeros(1, 4)]);
```

```
% Array mit wahren Schwellwerten
RealArray = [-3:0.25:3];

for i = 1:length(RealArray):1300
    for j = 1:length(RealArray)
        RealThrArray(i-1+j) = RealArray(j);
    end
end

% Main program
%--------------

for n=1:NumRuns

    if mod(n,2) == 0
        RealThr = RealThrArray(n);
    else
        RealThr = RealArray(length(RealArray)-(mod(n,length(RealArray))));
    end

    StepSize = InitialStepSize;
    trial = 1;
    M = zeros(NumTrials,1);
    R = zeros(NumTrials,1);
    M(1) = 0;

    Direction = 0;
    Reverse = 0;
    CountReversal = 0;

    while trial <= NumTrials

        % get response: R = 1 for yes, R = -1 for no
        %-------------------------------------------
        % psychometric function of observer
        RealProb = pg + ((1-pg-pl)/(1+exp(-(M(trial)-RealThr)/RealSlope)));
        ResponseArray = [ones(round(RealProb*100),1); ...
         -1*ones(round((1-RealProb)*100),1)];
        Randpoint = round(rand(1)*(length(ResponseArray)-1));
        R(trial) = ResponseArray(Randpoint+1);

        % next stimulus placement
        %------------------------

        % increase stimulus level after negative response
        if R(trial) == -1
            if Direction == -1
                Reverse = 1;
                CountReversal = CountReversal +1;
                if CountReversal == 1
                    StepSize = StepSize / 2;
                end
```

```
        else
            Reverse = 0;
        end

        M(trial+1) = M(trial) + StepSize;

        if M(trial+1) > maxRange
            M(trial+1) = maxRange;
        end
        Direction = 1;

    % decrease stimulus level after positive response
    elseif R(trial) == 1
        if Direction == 1
            Reverse = 1;
            CountReversal = CountReversal +1;
            if CountReversal == 1
                StepSize = StepSize / 2;
            end
        else
            Reverse = 0;
        end
        M(trial+1) = M(trial) - StepSize;

        if M(trial+1) < -maxRange
            M(trial+1) = -maxRange;
        end
        Direction = -1;
    end

    trial = trial + 1;

end %of all trials --> one threshold estimation run

% calculating final estimate
%---------------------------
Reversal = zeros(NumTrials,1);
NumTurn = 0;
Turnaround = 0;

for i=2:NumTrials

    % count reversal when response changes
    % save turnaround points
    if R(i) ~= R(i-1)
        Reversal(i) = 1;
        NumTurn = NumTurn + 1;
        Turnaround(NumTurn) = M(i);
    else
        Reversal(i) = 0;
    end
end
```

```
    % final estimate as average of turnaroundpoints
    EstimateThr = mean(Turnaround(StartMean:length(Turnaround)));
    EstimateThr = 0.25*round(EstimateThr*4);

    %Save data of current run
    allM(1:NumTrials,n) = M(1:NumTrials);
    allR(1:NumTrials,n) = R(1:NumTrials);
    allReversal(1:NumTrials,n) = Reversal(1:NumTrials);
    allTurnaround(1:length(Turnaround),n) = Turnaround;
    allRealThr(n) = RealThr;
    allEstimateThr(n) = EstimateThr;

end % of all threshold estimate runs

MeanEstimThr = mean(allEstimateThr);

% Speicherung der Daten
%----------------------

% Zuweisen der Werte-Vektoren auf Plätze im Array
Data.Stimuli(1:NumTrials,1:NumRuns) = allM(1:NumTrials,1:NumRuns);
Data.Responses(1:NumTrials,1:NumRuns) = allR(1:NumTrials,1:NumRuns);
Data.RealThresholds(1:NumRuns) = allRealThr(1:NumRuns);
Data.EstimateThresholds(1:NumRuns) = allEstimateThr(1:NumRuns);
Data.Sonstiges(10) = MeanEstimThr;

% Speichern des Arrays als mat-file
save(strcat('Simulation_Method', num2str(Methode), '_Paradigm', num2str(Para),
'_Trials', ... num2str(NumTrials), '_', '.mat'), 'Data');

% ========================================================================= %
```

B.2 Implementierung von PEST

```
% ========================================================================= %
% PEST Simulation nach Taylor & Creelman
% ========================================================================= %
clc; clear all; close all;

% Initialization
%----------------

% PEST = Methode 2
Methode = 2;
% Yes-No = 1, 2AFC = 2, 3AFC = 3
Para = 1;

% Anzahl der Durchläufe, Range und Trialzahl
NumRuns = 1000;
maxRange = 5;
NumTrials = 10;

pg = 0.03;                              % guessing rate
pl = 0.02;     % lapsing rate
InitialStepSize = 4;              % should be power of 2
TargetProb = (1 + pg - pl) / 2;  % target probability
W = 1;                            % deviation limit
FinalStepSize = 1/8;              % should be a power of 2
RealSlope = 2*maxRange/10;  % slope of psychometric function

% Definieren des leeren Daten-Arrays und der Parameter
Data = struct(  'Method', Methode, ...
                'Paradigm', Para,  ...
                'Trials', NumTrials,  ...
                'Stimuli', zeros(NumTrials,NumRuns), ...
                'Responses', zeros(NumTrials,NumRuns), ...
                'RealThresholds', zeros(1,NumRuns), ...
                'EstimateThresholds', zeros(1,NumRuns), ...
                'Sonstiges', [maxRange, pg, pl, RealSlope, InitialStepSize, ...
                 W, FinalStepSize, zeros(1,3)]);

% Array mit wahren Schwellwerten
RealArray = [-3:0.25:3];

for i = 1:length(RealArray):1300
    for j = 1:length(RealArray)
        RealThrArray(i-1+j) = RealArray(j);
    end
end
```

```
% Main program
%--------------

for n=1:NumRuns

    if mod(n,2) == 0
        RealThr = RealThrArray(n);
    else
        RealThr = RealArray(length(RealArray)-(mod(n,length(RealArray))));
    end

    trial = 1;
    CountTrial = 0;
    CountCorrect = 0;
    StepDirection = 0;
    SameDirectionStep = 0;
    Doubling1 = 0;
    Doubling2 = 0;
    NumReversal = 0;
    UnderStepSize = 0;

    M(1) = 0;                    % initial stimulus value
    StepSize = InitialStepSize; % initial step size

    while trial <= NumTrials

        % get response: R = 1 for correct, R = -1 for incorrect
        %------------------------------------------------------
        % psychometric function of observer
        RealProb = pg + ((1-pg-pl)/(1+exp(-(M(trial)-RealThr)/RealSlope)));
        ResponseArray = [ones(round(RealProb*100),1); ...
         -1*ones(round((1-RealProb)*100),1)];
        Randpoint = round(rand(1)*(length(ResponseArray)-1));
        R(trial) = ResponseArray(Randpoint+1);

        % Wald sequential likelihood-ratio test
        %--------------------------------------
        % Count trials and correct responses at same level
        CountTrial = CountTrial + 1;

        if R(trial) == 1
            CountCorrect = CountCorrect + 1;
        end

        % calculate lower and upper bound
        LowerBound = (TargetProb * CountTrial) - W;
        UpperBound = (TargetProb * CountTrial) + W;

        % current level is too high --> decrease level
        if CountCorrect >= UpperBound
            ChangeLevel = -1;
```

```
% current level is too low --> increase level
elseif CountCorrect <= LowerBound
    ChangeLevel = 1;
% next trial at same level
else
    ChangeLevel = 0;
end

% look for reversals and count steps in a given direction
%-------------------------------------------------------

% first case of reversal (from up to down)
if ChangeLevel == -1 && StepDirection == 1
    Reversal = 1;
    NumReversal = NumReversal + 1;
    SameDirectionStep = 1;
% second case of reversal (from down to up)
elseif ChangeLevel == 1 && StepDirection == -1
    Reversal = 1;
    NumReversal = NumReversal + 1;
    SameDirectionStep = 1;
% count another step in a given direction
elseif ChangeLevel ~= 0 && ChangeLevel == StepDirection
    Reversal = 0;
    SameDirectionStep = SameDirectionStep + 1;
% remain at current level
else
    Reversal = 0;
end


% next stimulus placement
%------------------------

% change stimulus level
if ChangeLevel ~= 0

    % halve step size on every reversal of step direction
    if Reversal == 1
        % halve step size
        StepSize = StepSize / 2;
        if StepSize == FinalStepSize
            StepSize = 2*FinalStepSize;
            UnderStepSize = UnderStepSize + 1;
        end
        % save state of doubling
        Doubling2 = Doubling1;
        Doubling1 = 0;
    % second step in same direction is same size as first
    elseif Reversal == 0 && SameDirectionStep == 2
        StepSize = StepSize;
```

```
            % third step size depending on sequence of preceding steps
            elseif Reversal == 0 && SameDirectionStep == 3
                if Doubling2 == 1
                    StepSize = StepSize;
                    Doubling1 = 0;
                elseif Doubling2 == 0
                    NumReversal = NumReversal - 1;
                    StepSize = StepSize * 2;
                    Doubling1 = 1;
                end
            % fourth and subsequent steps in given direction are doubled
            elseif Reversal == 0 && SameDirectionStep >= 4
                NumReversal = NumReversal - 1;
                StepSize = StepSize * 2;
                Doubling1 = 1;
            end

            M(trial+1) = M(trial) + (ChangeLevel * StepSize);

            % if exceeding upper and lower limit of range
            if M(trial+1) > maxRange
                M(trial+1) = maxRange;
            elseif M(trial+1) < -maxRange
                M(trial+1) = -maxRange;
            end

            StepDirection = ChangeLevel;
            CountTrial = 0;
            CountCorrect = 0;

        % else ask same level again
        else
            M(trial+1) = M(trial);
        end

        trial = trial + 1;

    end %of all trials --> one threshold estimation run finished

    EstimateThr = M(trial);

    %Save data of current run
    allM(1:trial,n) = M(1:trial);
    allR(1:trial-1,n) = R(1:trial-1);
    allRealThr(n) = RealThr;
    allEstimateThr(n) = EstimateThr;
    allNumReversals(n) = NumReversal;
    allUnderStepSize(n) = UnderStepSize;

end % of all threshold estimate runs
```

```
MeanEstimThr = mean(allEstimateThr);
MeanReversals = mean(allNumReversals);
MeanUnder = mean(allUnderStepSize);

% Speicherung der Daten
%----------------------

% Zuweisen der Werte-Vektoren auf Plätze im Array
Data.Stimuli(1:NumTrials,1:NumRuns) = allM(1:NumTrials,1:NumRuns);
Data.Responses(1:NumTrials,1:NumRuns) = allR(1:NumTrials,1:NumRuns);
Data.RealThresholds(1:NumRuns) = allRealThr(1:NumRuns);
Data.EstimateThresholds(1:NumRuns) = allEstimateThr(1:NumRuns);
Data.Sonstiges(8) = MeanReversals;
Data.Sonstiges(9) = MeanUnder;
Data.Sonstiges(10) = MeanEstimThr;

% Speichern des Arrays als mat-file
save(strcat('Simulation_Method', num2str(Methode), '_Paradigm', num2str(Para),
'_Trials', num2str(NumTrials), '_', '.mat'), 'Data');

% ========================================================================= %
```

B.3 Implementierung von Best PEST

```
% ========================================================================= %
% Best PEST Simulation nach Pentland
% ========================================================================= %
clc; clear all; close all;

% Initialization
%----------------

% Best PEST = Methode 3
Methode = 3;
% Yes-No = 1, 2AFC = 2, 3AFC = 3
Para = 1;

% Anzahl der Durchläufe, Range und Anzahl der Stimuli
NumRuns = 1000;
maxRange = 5;
NumTrials = 10;

pg = 0.03;  % guessing rate
pl = 0.02;  % lapsing rate

StimArray = [-5:0.25:5];
IndexRange = length(StimArray);
IndexArray = [1:IndexRange];

% Wahl der Steigung
RealSlope = 2*maxRange/10;
IndexSlope = (IndexRange-1)/10;

% Definieren des leeren Daten-Arrays und der Parameter
Data = struct(  'Method', Methode, ...
                'Paradigm', Para,  ...
                'Trials', NumTrials,  ...
                'Stimuli', zeros(NumTrials,NumRuns), ...
                'Responses', zeros(NumTrials,NumRuns), ...
                'RealThresholds', zeros(1,NumRuns), ...
                'EstimateThresholds', zeros(1,NumRuns), ...
                'Sonstiges', [maxRange, pg, pl, Slope, RealSlope, zeros(1, 5)]);


%Array mit wahren Schwellwerten
RealArray = [-3:0.25:3];

for i = 1:length(RealArray):1300
    for j = 1:length(RealArray)
        RealThrArray(i-1+j) = RealArray(j);
    end
end
```

```
% Main program
%--------------

for n=1:NumRuns

    if mod(n,2) == 0
        RealThr = RealThrArray(n);
    else
        RealThr = RealArray(length(RealArray)-(mod(n,length(RealArray))));
    end

    M = 0;
    R = 0;

    % implizite Trials
    if pg < 0.3
        M(1) = IndexRange; R(1) = 1;
        M(2) = 1; R(2) = -1;
        M(3) = round(IndexRange*0.5);
    else % auf 2AFC oder 3AFC anpassen
        M(1) = 1; R(1) = -1;
        M(2) = 1; R(2) = 1;
        M(3) = 1; R(3) = -1;
        M(4) = round(IndexRange*0.5);
    end

    % Initialisierung der Likelihood-Funktion
    Like = ones(1, IndexRange);

    for trial = 3:NumTrials

        % get response: R = 1 for yes, R = -1 for no
        %-------------------------------------------
        % psychometric function of observer
        RealProb = pg + ((1-pg-pl)/(1+exp(-(StimArray(M(trial))-RealThr)/RealSlope)));
        ResponseArray = [ones(round(RealProb*100),1); ...
         -1*ones(round((1-RealProb)*100),1)];
        Randpoint = round(rand(1)*(length(ResponseArray)-1));
        R(trial) = ResponseArray(Randpoint+1);

        % calculate maximum likelihood estimate in indices
        %-------------------------------------------------

        % fuer jeden bisher praesentierten Stimulus
        for i = 1:length(M)
            % fuer jede moegliche Stimulusstufe aus dem Wertebereich [1, range]
            for x = 1:IndexRange
                % Auftretenswahrscheinlichkeit einer positiven Antwort
                ProbPos = pg + ((1-pg-pl)/(1+exp(-(M(i)-x)/IndexSlope)));
                % Auftretenswahrscheinlichkeit einer negativen Antwort
                ProbNeg = 1 - ProbPos;
                if R(i) == 1  %falls positive Antwort
                    Like(x) = Like(x) * ProbPos;
```

```
                else    %falls negative Antwort
                    Like(x) = Like(x) * ProbNeg;
                end
          end
            % Normierung
            Like = Like / sum(Like);
        end

        % waehle Stimulus/-i im Maximum der Likelihood-Funktion
        val = find(Like==(max(Like)));
        % waehle einen Stimulus bei Gleichwahrscheinlichkeit mehrerer Alternativen
        val = round(mean(val));

        M(trial+1) = val;

    end % of all trials --> one threshold estimation run

    %Save data of current run
    allM(1:NumTrials,n) = StimArray(M(1:NumTrials));
    allR(1:NumTrials,n) = R(1:NumTrials);
    allRealThr(n) = RealThr;
    allEstimateThr(n) = StimArray(val);

end % of all threshold estimate runs

MeanEstimThr = mean(allEstimateThr);

% Speicherung der Daten
%----------------------

% Zuweisen der Werte-Vektoren auf Plätze im Array
Data.Stimuli(1:NumTrials,1:NumRuns) = allM(1:NumTrials,1:NumRuns);
Data.Responses(1:NumTrials,1:NumRuns) = allR(1:NumTrials,1:NumRuns);
Data.RealThresholds(1:NumRuns) = allRealThr(1:NumRuns);
Data.EstimateThresholds(1:NumRuns) = allEstimateThr(1:NumRuns);
Data.Sonstiges(10) = MeanEstimThr;

% Speichern des Arrays als mat-file
save(strcat('Simulation_Method', num2str(Methode), '_Paradigm', num2str(Para),
'_Trials', num2str(NumTrials), '_', '.mat'), 'Data');

% ========================================================================= %
```

B.4 Implementierung von ZEST

```
% ========================================================================= %
% ZEST Simulation nach King-Smith et al.
% ========================================================================= %
clc; clear all; close all;

% Initialization
%----------------

% ZEST = Methode 4
Methode = 4;
% Yes-No = 1, 2AFC = 2, 3AFC = 3
Para = 1;

% Anzahl der Durchläufe, Range und Trialzahl
NumRuns = 1000;
maxRange = 5;
NumTrials = 10;

StimArray = [-maxRange:0.25:maxRange];
IndexRange = length(StimArray);
IndexArray = [1:IndexRange];

pg = 0.03;      % guessing rate
pl = 0.02;       % lapsing rate

RealSlope = (2*maxRange)/10;
IndexSlope = (IndexRange-1)/10;
Std0 = 12;      % standard deviation of prior pdf

% Definieren des leeren Daten-Arrays und der Parameter
Data = struct(  'Method', Methode, ...
                'Paradigm', Para,  ...
                'Trials', NumTrials,  ...
                'Stimuli', zeros(NumTrials,NumRuns), ...
                'Responses', zeros(NumTrials,NumRuns), ...
                'RealThresholds', zeros(1,NumRuns), ...
                'EstimateThresholds', zeros(1,NumRuns), ...
                'Sonstiges', [maxRange, pg, pl, Indexbeta, Realbeta, StdQ0, zeros(1, 4)]);

% Array mit wahren Schwellwerten
RealArray = [-3:0.25:3];
RealIndex = [-12:12];

for i = 1:length(RealArray):1300
    for j = 1:length(RealArray)
        RealThrArray(i-1+j) = RealIndex(j);
    end
end
```

```
% Main program
%--------------

for n=1:NumRuns

    if mod(n,2) == 0
        RealThr = RealThrArray(n); % variable real threshold
    else
        RealThr = RealIndex(length(RealIndex)-(mod(n,length(RealIndex))));
    end

    % Berechnung der psychometrischen Funktionen
    for i = 1:2*IndexRange
        PsyFunc(i) = pg + ((1-pg-pl)/(1+exp(-(i - IndexRange)/IndexSlope)));
    end

    for i = 1:2*IndexRange-1
        %Failure(i) = 1-PsyFunc(-i);
        ResFunc(1, i) = (1-PsyFunc(2*IndexRange-i));
        %Success(i) = PsyFunc(-i);
        ResFunc(2, i) = (PsyFunc(2*IndexRange-i));
    end

    % initialize prior pdf and initial pdf function
    for t = 1:IndexRange
        Zest0(t) = -(((t-IndexRange/2)/Std0)^2)+18;
    end

    Zest0 = Zest0 / sum(Zest0);
    Zest = Zest0;

    % begin of trials
    for trial = 1:NumTrials

        % next stimulus placement
        % -----------------------
        MeanStim = 0;

        for k = 1:IndexRange
            MeanStim = MeanStim + Zest(k) * IndexArray(k);
        end

        x(trial) = round(MeanStim);

        % get response: R = 2 for yes, R = 1 for no
        %------------------------------------------
        % psychometric function of observer
        RealProb = pg + ((1-pg-pl)/(1+exp(-(x(trial) - (round(IndexRange/2) + RealThr))/Ind
        ResponseArray = [2*ones(round(RealProb*100),1); ...
         ones(round((1-RealProb)*100),1)];
        Randpoint = round(rand(1)*(length(ResponseArray)-1));
        R(trial) = ResponseArray(Randpoint+1);
```

```
        % calculate ZEST function
        % ------------------------
        for t = 1:IndexRange
            Zest(t) = Zest(t) * ResFunc(R(trial), IndexRange + t - x(trial));
        end

        % Normierung
        Zest = Zest / sum(Zest);

    end % of all trials

    % calculate final estimate
    % ------------------------
    % waehle den (oder die) Stimulus/-i im Mean der Likelihood-Funktion
    FinalMean = 0;

    for k = 1:IndexRange
        FinalMean = FinalMean + Zest(k) * IndexArray(k);
    end

    %Save data of current run
    allM(1:NumTrials,n) = x(1:NumTrials);
    allR(1:NumTrials,n) = R(1:NumTrials);
    allRealThr(n) = StimArray(round(IndexRange/2) + RealThr);
    allEstimateThr(n) = StimArray(round(FinalMean));

end % of all threshold estimate runs

MeanEstimThr = mean(allEstimateThr);

% Speicherung der Daten
%-----------------------

% Zuweisen der Werte-Vektoren auf Plätze im Array
Data.Stimuli(1:NumTrials,1:NumRuns) = allM(1:NumTrials,1:NumRuns);
Data.Responses(1:NumTrials,1:NumRuns) = allR(1:NumTrials,1:NumRuns);
Data.RealThresholds(1:NumRuns) = allRealThr(1:NumRuns);
Data.EstimateThresholds(1:NumRuns) = allEstimateThr(1:NumRuns);
Data.Sonstiges(10) = MeanEstimThr;

% Speichern des Arrays als mat-file
save(strcat('Simulation_Method', num2str(Methode), '_Paradigm', num2str(Para),
'_Trials', num2str(NumTrials), '_.mat'), 'Data');

% ========================================================================= %
```

B.5 Auswertung der Ergebnisse

```
% ========================================================================= %
% Auswertung der Simulationsdaten
%=========================================================================%
clc; clear all; close all;


% Einlesen der Daten aus mat-files und Ergebnis-Matrix erstellen
% ---------------------------------------------------------------

% 1000 Messreihen (Stimuliplatzierung) & 1000 Antwort-Arrays
% 1000 reale Schwellwerte & 1000 ermittelte Schwellwerte
% über 4 Methoden & 3 Paradigmen & 5 Triallängen
DATA = zeros(1000, 2, 4, 3, 5);

% 4 Methoden
for i=1:4
    % 3 Paradigmen
    for j=1:3
        % 5 Triallängen
        for k=1:5
            name = strcat('/Users/stefcia/Documents/kommunikation/Magisterarbeit/...
            Simulation in Matlab/Auswertung/Simulation Data neu/ Simulation_Method',
            num2str(i), '_Paradigm',num2str(j),'_Trials',num2str(k*10),'_');
            load (name)
            DATA(1:length(Data.RealThresholds), 1, i, j, k) = Data.RealThresholds;
            DATA(1:length(Data.EstimateThresholds), 2, i, j, k) = Data.EstimateThresholds;
        end
    end
end

% Varianz, Bias und Effizienz berechnen
% -------------------------------------

% Mittelwert der realen Schwelle
MeanReal = mean(DATA(:,1,1,1,1));
% Standardabweichung der realen Schwelle
StdReal = std(DATA(:,1,1,1,1));

% 4 Methoden
for i=1:4
    % 3 Paradigmen
    for j=1:3
        % 5 Triallängen
        for k=1:5
            % Mittelwert der ermittelten Schwelle
            MeanThres(i,j,k) = mean(DATA(:,2,i,j,k));
            % Standardabweichung der ermittelten Schwelle
            StdThres(i,j,k) = std(DATA(:,2,i,j,k));
            % Varianz
            Var(i,j,k) = (StdThres(i,j,k))^2;
```

```
            % Effizienz bzw. sweat factor
            sweatfactor(i,j,k) = k*10 * Var(i,j,k);
            % Fehler der ermittelten Schwellwerte (real - estimate)
            Error(:,i,j,k) = DATA(:,2,i,j,k) - DATA(:,1,i,j,k);
            % Bias
            Bias(i,j,k) = mean(Error(:,i,j,k));
            % Varianz des Bias
            VarBias(i,j,k) = (std(Error(:,i,j,k)))^2;
        end
    end
end

% ========================================================================== %
```