

The logo for 'esz' features the lowercase letters 'esz' in a bold, black, sans-serif font. The letters are centered within a white square that is partially enclosed by a yellow border. Two yellow rectangular blocks are positioned above and below the 's', with black curved lines connecting them to the top and bottom of the 's' respectively.

**esz**

**calibration & metrology**



# EIGNUNGSPRÜFUNGEN ZWISCHEN KALIBRIERLABORATORIEN

**STATISTISCHE AUSWERTUNG VON VERGLEICHSMESSUNGEN**



## White Paper

Titel: Eignungsprüfungen zwischen Kalibrierlaboratorien –  
Statistische Auswertung von Vergleichsmessungen

Version: 1.0

Ausgabedatum: 24. Oktober 2024

Autorin: Patricia Mayer

Kontakt: [pt@esz-ag.de](mailto:pt@esz-ag.de)

© esz AG calibration & metrology

# Inhaltsverzeichnis

<b>1</b>	<b>Einleitung</b>	<b>1</b>
<b>2</b>	<b>Theoretische Grundlagen</b>	<b>2</b>
2.1	Datensatz . . . . .	2
2.2	Kriterium zur Leistungsbewertung — der $E_n$ -Wert . . . . .	3
2.2.1	Graphische Interpretation des $E_n$ -Wertes. . . . .	5
2.2.2	Leistungsbewertung bei $u_i < u_{ref}$ . . . . .	7
2.2.3	Der $E_n$ -Wert als Hypothesentest . . . . .	8
2.2.4	Berücksichtigung von Korrelationen im $E_n$ -Wert . . . . .	13
<b>3</b>	<b>Methode A — Referenzlabor</b>	<b>17</b>
<b>4</b>	<b>Methode B — Gewichteter Mittelwert (Cox)</b>	<b>20</b>
4.1	Berechnung der Referenzdaten . . . . .	20
4.1.1	Chi-Quadrat-Test . . . . .	22
4.1.2	Leistungsbewertung . . . . .	22
4.2	Charakteristik der Methode . . . . .	23
4.2.1	Eigenschaften der Referenzunsicherheit . . . . .	23
4.2.2	Eigenschaften des Referenzwertes . . . . .	25
4.2.3	Eigenschaften der Leistungsbewertung . . . . .	28
<b>5</b>	<b>Methode C — Robuste Abwandlung des arithmetischen Mittelwertes (ISO 13528:2022)</b>	<b>34</b>
5.1	Arithmetischer Mittelwert (Methode $C_0$ ) . . . . .	34
5.1.1	Charakteristik der Methode . . . . .	37
5.2	Robuster Algorithmus (Methode C) . . . . .	40
5.2.1	Berechnung der Referenzdaten . . . . .	40
5.2.2	Leistungsbewertung . . . . .	46
5.2.3	Charakteristik der Methode . . . . .	49
<b>6</b>	<b>Methode D — Potenz-moderierter Mittelwert (Pommé &amp; Keightley)</b>	<b>58</b>
6.1	Mandel-Paule Mittelwert (Methode $D_0$ ) . . . . .	58
6.1.1	Berechnung der Referenzdaten . . . . .	58
6.1.2	Leistungsbewertung . . . . .	63
6.1.3	Charakteristik der Methode . . . . .	64
6.2	Potenz-moderierter Mittelwert (Methode D) . . . . .	70
6.2.1	Berechnung der Referenzdaten . . . . .	70
6.2.2	Leistungsbewertung und Ausschluss von Teilnehmerdaten . . . . .	74
6.2.3	Charakteristik der Methode . . . . .	76
<b>7</b>	<b>Vergleich der Auswertungsmethoden</b>	<b>81</b>
7.1	Beispiel 1: Konsistenter Datensatz . . . . .	82
7.2	Beispiel 2: Datensatz mit Ausreißern . . . . .	83

7.3	Beispiel 3: Ein Teilnehmer mit kleiner Unsicherheit . . . . .	84
7.4	Beispiel 4: Geringe Streuung; große, ähnliche Messunsicherheit . . . . .	85
7.5	Beispiel 5: Mäßig konsistenter Datensatz . . . . .	86
7.6	Beispiel 6: Viele Teilnehmer . . . . .	87
7.7	Beispiel 7: Wenig Teilnehmer . . . . .	88
<b>8</b>	<b>Bilaterale Eignungsprüfungen</b>	<b>89</b>
8.1	Methode B . . . . .	89
8.2	Methode $C_0$ . . . . .	91
8.3	Methode C . . . . .	91
8.4	Methode $D_0$ . . . . .	92
8.5	Methode D . . . . .	93
	<b>Abbildungsverzeichnis</b>	<b>94</b>
	<b>Literaturverzeichnis</b>	<b>96</b>

# 1 Einleitung

Gemäß DIN EN ISO/IEC 17025:2018 [1], Absatz 7.7.2, ist ein akkreditiertes Kalibrierlabor dazu verpflichtet, „sofern verfügbar und zweckmäßig, seine Leistung durch Vergleich mit den Ergebnissen anderer Laboratorien [zu] überwachen“. Eine Möglichkeit dies zu tun, ist nach Absatz 7.7.2.a) durch die Teilnahme an einer Eignungsprüfung gegeben. Die Organisation und Auswertung einer Eignungsprüfung übernimmt ein nach DIN EN ISO/IEC 17043:2023 [2] akkreditierter Eignungsprüfungsanbieter. Das vorliegende Dokument behandelt die statistische Auswertung von Eignungsprüfungen. In diesem Zusammenhang verweist DIN EN ISO/IEC 17043:2023 auf ISO 13528:2022 [3]. Auf die Organisation und Durchführung einer Eignungsprüfung wird in diesem Dokument lediglich im Rahmen der benötigten Informationen eingegangen. Gleiches gilt für etwaige Stabilitätsbetrachtungen des Eignungsprüfungsgegenstandes, d.h. es wird von einem idealen, stabilen Eignungsprüfungsgegenstand ausgegangen.

Der Datensatz der Eignungsprüfung und das Kriterium zur Leistungsbewertung, der  $E_{n;i}$ -Wert, werden in Kapitel 2 definiert. Für letzteren werden sowohl graphische, als auch statistische Eigenschaften aufgezeigt und analysiert. Ebenso wird auf die Berücksichtigung von Korrelationen eingegangen.

Für die Bewertung der Teilnehmerleistung über den  $E_{n;i}$ -Wert werden Referenzdaten benötigt, welche bspw. durch die Messungen eines Referenzlabors festgelegt werden können, wie Kapitel 3 zeigt. Diese Vorgehensweise wird in diesem Dokument als Methode A bezeichnet.

Der Fokus dieses Dokumentes liegt auf einer statistischen Bestimmung der benötigten Referenzdaten unter Einbeziehung von Teilnehmerdaten. Kapitel 4 betrachtet den von Maurice Cox in [4] beschriebenen gewichteten Mittelwert, der im Rahmen dieses Dokumentes als Methode B bezeichnet wird. Hierbei handelt es sich um eine unsicherheitsbasierte Auswertungsmethode mit individueller Teilnehmergewichtung. Dem gegenüber stehen die Methoden  $C_0$  und C, bei denen es sich um den arithmetischen Mittelwert und um eine nach [3] robuste Abwandlung desselben handelt. Charakteristisch ist hier, dass es sich um eine streuungsbasierte Auswertung mit einheitlicher Teilnehmergewichtung handelt. Als Kompromiss zwischen den Methoden B und  $C_0/C$  werden in Kapitel 6 der Mandel-Paul Mittelwert nach [5, 6] als Methode  $D_0$  und der potenz-moderierte Mittelwert aus [7, 8] als Methode D eingeführt. Beide Methoden können zwischen den Grenzfällen von Methode B und Methode  $C_0/C$  moderieren, indem sowohl die Teilnehmerunsicherheiten, als auch die Streuung der Teilnehmerwerte berücksichtigt werden. Durch die Möglichkeit verschiedene Grenz- und Zwischenfälle abzubilden und die Anwendbarkeit bei einer Vielzahl an Datensätzen, eignen sich die Methoden  $D_0$  und D als Standardauswertungsmethoden. Ziel dieses Dokumentes ist es, die genannten Methoden vorzustellen, sowie deren Charakteristik zu analysieren und zu vergleichen.

Anhand ausgewählter Beispieldaten werden die Eigenschaften der genannten Auswertungsmethoden in Kapitel 7 dargestellt. Abschließend betrachtet Kapitel 8 den Grenzfall einer bilateralen Eignungsprüfung und zeigt auf, dass diese letztlich nur durch Methode A sinnvoll ausgewertet werden kann.

## 2 Theoretische Grundlagen

Zunächst werden in diesem Kapitel die benötigten Grundlagen einer Eignungsprüfung definiert, dargestellt und analysiert. Dazu gehören sowohl der Datensatz einer Eignungsprüfung, als auch das Kriterium zur Leistungsbewertung. Hierfür wird im Rahmen dieses Dokumentes von einem idealen, d.h. stabilen Eignungsprüfungsgegenstand ausgegangen. Stabilitätsuntersuchungen, wie sie von der DIN EN ISO/IEC 17043:2023 [2] vorgeschrieben werden, sind nicht Gegenstand der nachfolgenden Betrachtungen.

### 2.1 Datensatz

Es sei  $N$  die Gesamtzahl der an der Eignungsprüfung teilnehmenden Labore und  $X$  die interessierende Messgröße der Eignungsprüfung. Vom Eignungsprüfungsanbieter wird ein für diese Messgröße geeigneter Eignungsprüfungsgegenstand zur Verfügung gestellt. Der Datensatz, der nach Durchführung aller Messungen vorliegt, setzt sich zusammen aus

- dem Kalibrierwert  $x_{i;0}$  (auch Nominalwert; nach VIM:2007 Absatz 5.18 [9] wird hierfür auch der Begriff Referenzwert verwendet. Dieser Definition wird in diesem Dokument nicht gefolgt um eine Vermischung mit dem Referenzwert der Eignungsprüfung zu vermeiden.),
- dem Messwert  $x_{i;\text{meas}}$  (s. VIM:2007 Absatz 2.10 [9])
- und der dazugehörigen erweiterten Messunsicherheit  $U_i$  (nach VIM:2007 Absatz 2.26 [9])

aller Teilnehmer  $i$  mit  $i \in [1; N]$  zur Messgröße  $X$ . Für die Auswertung wird die von Teilnehmer  $i$  festgestellte Messabweichung  $x_i$  nach VIM:2007 Absatz 2.16 [9] betrachtet

$$x_i = x_{i;\text{meas}} - x_{i;0}. \quad (2.1)$$

Das Messergebnis eines Teilnehmers  $i \in [1; N]$  ist somit gegeben durch

$$x_i \pm U_i. \quad (2.2)$$

Die erweiterte Messunsicherheit  $U_i$  des Teilnehmers  $i$  setzt sich zusammen aus dessen Erweiterungsfaktor  $k_i$  und der Standardunsicherheit  $u_i$

$$U_i = k_i \cdot u_i. \quad (2.3)$$

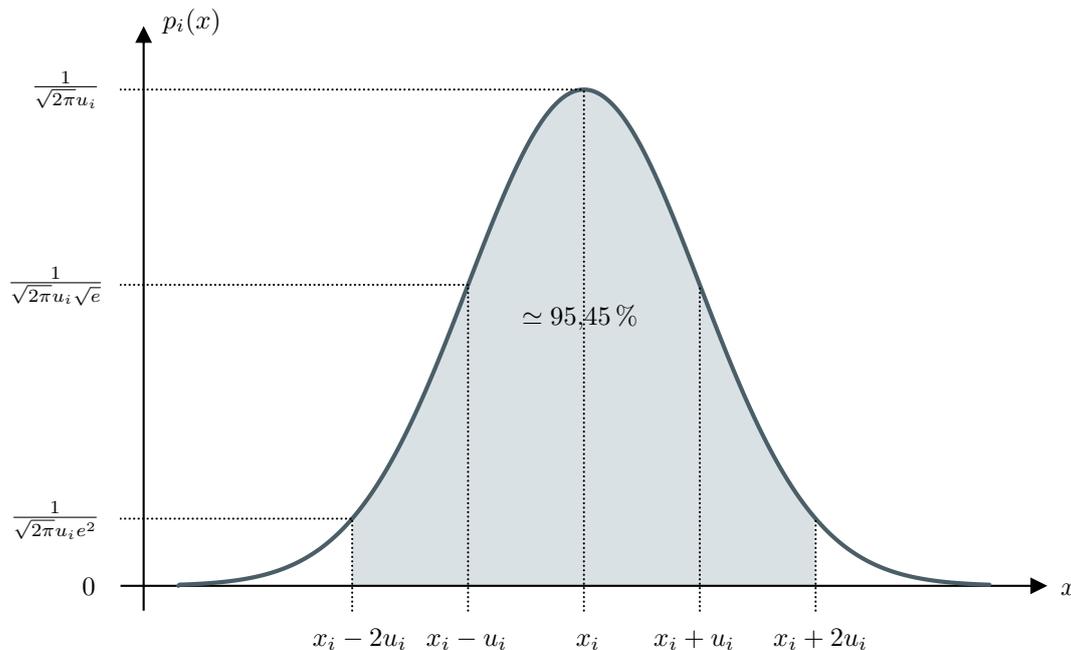
Es wird nachfolgend angenommen, dass das Messergebnis in Gleichung (2.2) für alle Teilnehmer  $i \in [1; N]$  durch eine Normalverteilung beschrieben wird. Unter dieser Annahme gilt  $k_i = 2 \forall i \in [1; N]$ , um eine Überdeckungswahrscheinlichkeit von ca. 95,45 % zu erzielen. Damit ist die erweiterte Messunsicherheit für alle Teilnehmer gegeben durch

$$U_i = 2 \cdot u_i \quad \forall i \in [1; N]. \quad (2.4)$$

Abweichungen von dieser Annahme sind gesondert zu betrachten und nicht Gegenstand dieses Dokumentes. Unter der getroffenen Annahme liegt allen Messergebnissen die in Abbildung 2.1 dargestellte Wahrscheinlichkeitsdichte zugrunde, die durch

$$p_i(x) = \frac{1}{\sqrt{2\pi}u_i} e^{-\frac{(x-x_i)^2}{2u_i^2}} \quad (2.5)$$

ausgedrückt werden kann. Der Erwartungswert ist folglich gegeben durch  $x_i$  und die Standardabweichung der Normalverteilung ist definiert durch  $u_i$ .



**Abbildung 2.1**

Es wird angenommen, dass die Messergebnisse aller Teilnehmer  $i \in [1; N]$  durch eine Normalverteilung mit Erwartungswert  $x_i$  und Standardabweichung  $u_i$  beschrieben werden. Der Bereich  $[x_i - 2u_i; x_i + 2u_i]$  entspricht einer Überdeckungswahrscheinlichkeit von ca. 95,45% und führt zur Verwendung des Erweiterungsfaktors  $k_i = 2$ , sodass die erweiterte Messunsicherheit gegeben ist durch  $U_i = 2 \cdot u_i$ .

Anmerkung: Im Rahmen dieses Dokuments wird angenommen, dass zwischen den einzelnen Teilnehmerdaten keine Korrelation vorliegt, sodass die Kovarianz zweier Teilnehmerwerte  $x_i$  und  $x_j$  null ist, d.h.  $u(x_i; x_j) = 0 \forall i \neq j \in [1; N]$ . Diese Annahme ist insbesondere dann zu überprüfen, wenn die Datensätze zweier Teilnehmer einem Labor oder einem Unternehmen zugeordnet werden können. Denkbar wäre zum Beispiel, dass ein Labor an einer Eignungsprüfung teilnimmt und zwei oder mehr Datensätze abgibt, die mit jeweils unterschiedlichen Verfahren, Prüfmitteln oder an verschiedenen Standorten bestimmt wurden. In diesem Fall kann eine Korrelation der Teilnehmerergebnisse nicht a-priori ausgeschlossen werden. Dieser Punkt ist insbesondere bei der Anwendung der Methoden B (4), C (5) und D (6) zu bedenken und idealerweise auszuschließen.

## 2.2 Kriterium zur Leistungsbewertung — der $E_n$ -Wert

Das Ziel der Eignungsprüfung ist es letztlich die Leistung eines Teilnehmers zu bewerten. Hierfür gibt es verschiedene Herangehensweisen. Das in diesem Dokument betrachtete Kriterium zur Leistungs-

bewertung ist durch den sogenannten  $E_n$ -Wert (normalized Error) gegeben, dessen Berechnung auf der Definition des „degree of equivalence“ in [10] basiert. Hierfür wird zunächst die Differenz  $d_i$  des Teilnehmerwertes  $x_i$  zu einem Referenzwert  $x_{\text{ref}}$  bestimmt

$$d_i = x_i - x_{\text{ref}}. \quad (2.6)$$

In Absatz 2.1 wurde die Annahme gemacht, dass das Ergebnis des Teilnehmers  $i$  durch eine Normalverteilung beschrieben werden kann. Diese Annahme wird nun auch für das Referenzergebnis  $x_{\text{ref}} \pm 2u_{\text{ref}}$  (d.h.  $U_{\text{ref}} = 2u_{\text{ref}}$ ) getroffen. Damit unterliegt auch die Differenz  $d_i$  aus Gleichung (2.6) einer Normalverteilung. Für die dazugehörige erweiterte Unsicherheit  $U_{d_i}$  gilt daher gemäß JCGM 100:2008 (GUM) [11] und der Vorgabe eines Vertrauensniveaus von ca. 95,45 % in [10]

$$\begin{aligned} U_{d_i} &= 2\sqrt{u_i^2 + u_{\text{ref}}^2 + 2\frac{\partial d_i}{\partial x_i} \frac{\partial d_i}{\partial x_{\text{ref}}} u(x_i; x_{\text{ref}})} \\ &= 2\sqrt{u_i^2 + u_{\text{ref}}^2 - 2u(x_i; x_{\text{ref}})}, \end{aligned} \quad (2.7)$$

d.h. es wird der Erweiterungsfaktor  $k = 2$  gewählt. Die Kovarianz  $u(x_i; x_{\text{ref}})$  ist zu berücksichtigen, falls eine Korrelation zwischen Teilnehmerwert  $x_i$  und Referenzwert  $x_{\text{ref}}$  vorliegt. Näheres dazu ist unter Abschnitt 2.2.4 zu finden. Die in Gleichung (2.6) gegebene Differenz wird auf ihre erweiterte Unsicherheit aus Gleichung (2.7) normiert und definiert so den  $E_{n;i}$ -Wert, die Variable für die Leistungsbewertung des Teilnehmers  $i$

$$\begin{aligned} E_{n;i} &= \frac{d_i}{U_{d_i}} \\ &= \frac{x_i - x_{\text{ref}}}{2\sqrt{u_i^2 + u_{\text{ref}}^2 - 2u(x_i; x_{\text{ref}})}} \end{aligned} \quad (2.8)$$

Das Kriterium für eine positive Bewertung der Leistung wird, entsprechend [10] so definiert, dass

$$-1 \leq E_{n;i} \leq 1. \quad (2.9)$$

Eine genauere Betrachtung diese Kriteriums zur Leistungsbewertung ist in Abschnitt 2.2.3 gegeben.

Anmerkung: Im Rahmen dieses Dokuments wird davon ausgegangen, dass ein idealer, d.h. im Rahmen der Eignungsprüfung stabiler, Eignungsprüfungsgegenstand vorliegt. Ist dies nicht der Fall, so ist die Auswertung und die Definition des  $E_{n;i}$ -Wertes aus Gleichung (2.8) ggf. durch einen Stabilitätsbeitrag zu erweitern. Alternativ existieren im Falle von Instabilitäten auch entsprechende Auswertungsmethoden zur Bestimmung der Referenzdaten [12]. Diese werden im Rahmen dieses Dokuments jedoch nicht betrachtet.

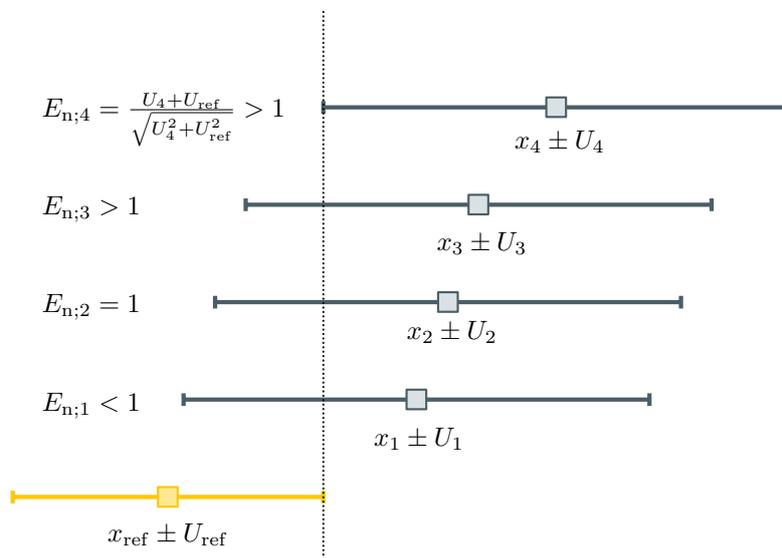
## 2.2.1 Graphische Interpretation des $E_n$ -Wertes.

Umgangssprachlich wird eine Eignungsprüfung oft als bestanden bezeichnet, wenn das betreffende Teilnehmerergebnis  $x_i \pm U_i$  im Rahmen der Messunsicherheit mit dem Referenzergebnis  $x_{\text{ref}} \pm U_{\text{ref}}$  übereinstimmt. Im Sinn der Ausführung in Abschnitt 2.2.3 ist diese Formulierung gültig. Allerdings darf dies nicht mit einer graphischen Überlappung der Unsicherheitsbalken verwechselt werden, da diese nicht mit dem in Gleichung (2.9) definierten Kriterium gleichzusetzen ist, wie Abbildung 2.2 zeigt. Die Überlappung der Unsicherheitsbalken ist demnach ein notwendiges, aber kein hinreichendes Kriterium für eine positive Leistungsbewertung über den  $E_{n;i}$ -Wert gemäß Gleichung (2.9). Würde die Überlappung der Unsicherheitsbalken als Kriterium für die Leistungsbewertung herangezogen, so müsste eine entsprechende Variable gemäß  $(x_i - x_{\text{ref}})/(U_i + U_{\text{ref}})$  definiert werden. Der Unterschied zu Gleichung (2.8) ist folglich die Addition der Unsicherheiten, wobei zu beachten ist, dass allgemein  $U_i + U_{\text{ref}} > \sqrt{U_i^2 + U_{\text{ref}}^2} \forall U_i, U_{\text{ref}} > 0$  (s. dazu bspw. [13]). Dadurch ist unter der Annahme, dass keine Korrelationen vorliegen klar, dass

$$E_{n;i} = 1 \Leftrightarrow x_i - x_{\text{ref}} = \sqrt{U_i^2 + U_{\text{ref}}^2} < U_i + U_{\text{ref}} \quad (2.10)$$

bzw.

$$E_{n;i} = \frac{x_i - x_{\text{ref}}}{\sqrt{U_i^2 + U_{\text{ref}}^2}} \stackrel{|x_i - x_{\text{ref}}| = U_{\text{ref}} + U_i}{=} \frac{U_i + U_{\text{ref}}}{\sqrt{U_i^2 + U_{\text{ref}}^2}} > 1. \quad (2.11)$$



**Abbildung 2.2**

Die Überlappung der Unsicherheitsbalken von Teilnehmerergebnis und Referenzergebnis ist nicht mit einer positiven  $E_{n;i}$ -Auswertung gleichzusetzen. Für die Graphik wurden folgende Werte verwendet:  $x_{\text{ref}} = 0$ ;  $x_1 = 1,6$ ;  $x_2 = \sqrt{1^2 + 1,5^2}$ ;  $x_3 = 2$ ;  $x_4 = 2,5$ ;  $U_{\text{ref}} = 1$ ;  $U_i = 1,5 \forall i \in [1; 4]$ . Es wurde angenommen, dass zwischen Teilnehmerwert und Referenzwert keine Korrelation vorliegt.

Wenn, wie in Abbildung 2.3 gezeigt, der Teilnehmerwert  $x_i$  innerhalb der Referenzunsicherheit liegt,

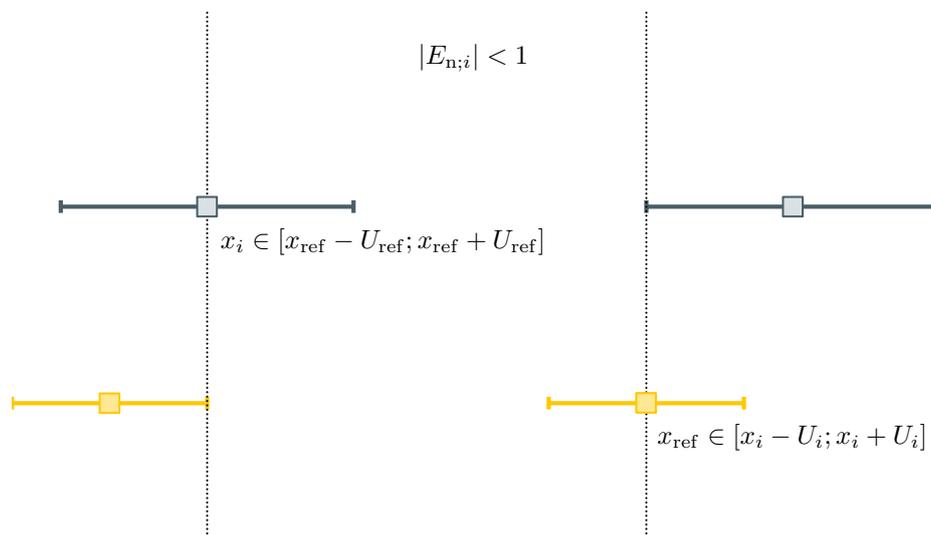
d.h.  $x_i \in [x_{\text{ref}} - U_{\text{ref}}; x_{\text{ref}} + U_{\text{ref}}]$  oder der Referenzwert innerhalb der Teilnehmerunsicherheit liegt, d.h.  $x_{\text{ref}} \in [x_i - U_i; x_i + U_i]$ , dann ist in beiden Fällen der  $E_{n;i}$ -Wert betragsmäßig kleiner als 1 und die Leistung des Teilnehmers  $i$  wird positiv bewertet, denn es gilt

$$|E_{n;i}| = \frac{|x_i - x_{\text{ref}}|}{\sqrt{U_i^2 + U_{\text{ref}}^2}} \leq \begin{cases} \frac{U_{\text{ref}}}{\sqrt{U_i^2 + U_{\text{ref}}^2}} & x_i \in [x_{\text{ref}} - U_{\text{ref}}; x_{\text{ref}} + U_{\text{ref}}] \\ \frac{U_i}{\sqrt{U_i^2 + U_{\text{ref}}^2}} & x_{\text{ref}} \in [x_i - U_i; x_i + U_i] \end{cases} < 1 \quad (2.12)$$

In diesem Spezialfall kann somit bereits anhand der graphischen Darstellung erkannt werden, dass die Leistung des Teilnehmers positiv zu bewerten ist. Gleichung (2.13) zeigt die mathematische Verallgemeinerung dieses hinreichenden Kriteriums für eine positive Leistungsbewertung.

$$|x_i - x_{\text{ref}}| \leq \max(U_i; U_{\text{ref}}) \Rightarrow |E_{n;i}| < 1 \quad (2.13)$$

An dieser Stelle sei noch angemerkt, dass die in Gleichung (2.8) und Gleichung (2.9) gezeigte Definition des Kriteriums zur Leistungsbewertung, den Vorteil bietet, dass die Information  $x_i \leq x_{\text{ref}}$  nicht verloren geht. Bei einer betragsmäßigen Definition des  $E_n$ -Wertes ist diese Information nicht mehr enthalten.



### Abbildung 2.3

Falls der Teilnehmerwert  $x_i$  innerhalb des Unsicherheitsintervalls des Referenzwertes liegt oder der Referenzwert innerhalb des Unsicherheitsintervalls des Teilnehmers, ist der  $E_{n;i}$ -Wert betragsmäßig kleiner als 1. Eine graphisch Abschätzung der Leistung ist in diesem Fall möglich.

## 2.2.2 Leistungsbewertung bei $u_i < u_{\text{ref}}$

In Analogie zu Gleichung (2.13) kann ein weiteres hinreichendes Kriterium für eine positive Leistungsbewertung durch das Minimum von Teilnehmer- und Referenzunsicherheit formuliert werden. Hierfür wird die Abschätzung  $\sqrt{U_i^2 + U_{\text{ref}}^2} \leq \sqrt{2} \min(U_i; U_{\text{ref}})$  verwendet, sodass folgt

$$|x_i - x_{\text{ref}}| \leq \sqrt{2} \min(U_i; U_{\text{ref}}) \Rightarrow |E_{n;i}| < 1. \quad (2.14)$$

Das Gleichheitszeichen in Gleichung (2.14) ist nur dann gültig, wenn  $U_i \neq U_{\text{ref}}$ .

Im Allgemeinen ist ein notwendiges und hinreichendes Kriterium für die positive Leistungsbewertung die Erfüllung von Gleichung (2.9). Zusätzlich wird gefordert, dass die Referenzunsicherheit  $u_{\text{ref}}$  kleiner ist als die Teilnehmerunsicherheit  $u_i$ , sodass in Gleichung (2.14) gilt, dass  $\min(U_i; U_{\text{ref}}) = U_{\text{ref}}$ . Auch wenn es für einen Teilnehmer stets das Ziel sein muss einer Eignungsprüfung teilzunehmen, deren Referenzunsicherheit kleiner ist als die eigene Teilnehmerunsicherheit, so können Fälle auftreten, in denen das nicht möglich oder praktikabel ist. Dies ist vor allem dann problematisch, wenn eine positive Leistungsbewertung gemäß Gleichung (2.9) nur durch die größere Referenzunsicherheit zustande kommt, d.h. wenn der Teilnehmer von der größeren Referenzunsicherheit profitiert. Um dies zu vermeiden, wird vorgeschlagen, dass in diesen Fällen Gleichung (2.14) als notwendiges Kriterium für eine positive Leistungsbewertung verwendet wird, d.h. die Erfüllung von

$$|x_i - x_{\text{ref}}| \leq \sqrt{2} U_i \quad (2.15)$$

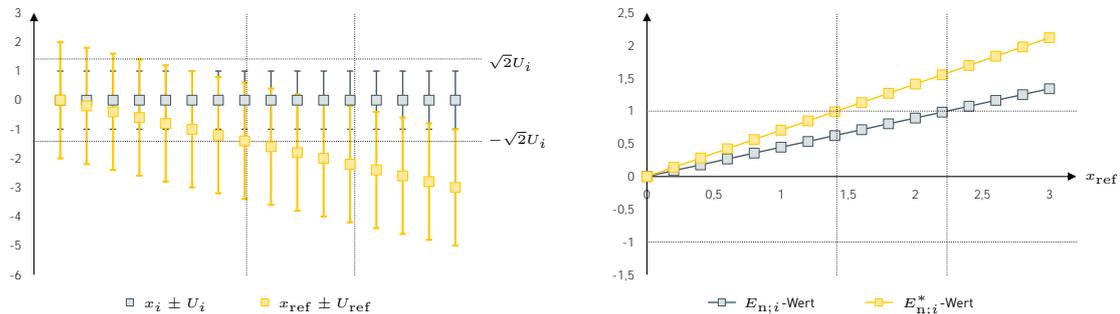
führt zu einer positiven Leistungsbewertung wenn  $u_i < u_{\text{ref}}$ . Voraussetzung für die Anwendbarkeit von Gleichung (2.15) ist, dass zwischen Teilnehmer- und Referenzwert keine Korrelation vorliegt. In dieser Definition der Leistungsbewertung stellt die Forderung  $|E_{n;i}| < 1$  ein notwendiges, aber kein hinreichendes Kriterium dar, sodass Gleichung (2.15) ein strengeres Kriterium darstellt als die Leistungsbewertung über Gleichung (2.9). Es sei angemerkt, dass Gleichung (2.15) einer modifizierten  $E_{n;i}^*$ -Auswertung gemäß Gleichung (2.16) entspricht, wobei ausgehend von Gleichung (2.8) die Referenzunsicherheit durch die Teilnehmerunsicherheit ersetzt wird. Die Kovarianz entfällt, da angenommen wird, dass Teilnehmer- und Referenzwert unkorreliert sind.

$$E_{n;i}^* = \frac{x_i - x_{\text{ref}}}{\sqrt{2} U_i} \quad (2.16)$$

Die Leistung wird positiv bewertet, wenn

$$-1 < E_{n;i}^* < 1. \quad (2.17)$$

Die vorgeschlagene Leistungsbewertung gemäß Gleichung (2.16) ist unabhängig von der Referenzunsicherheit, wodurch im Falle  $u_i < u_{\text{ref}}$  sicher gestellt wird, dass der Teilnehmer nicht von der größeren Referenzunsicherheit profitiert. Die Größe der Referenzunsicherheit legt letztlich nur noch fest, wie klein der  $E_{n;i}$ -Wert aus Gleichung (2.8) wird. Abbildung 2.4 zeigt einen Vergleich zwischen den Leistungsbewertungen gemäß Gleichung (2.8) und Gleichung (2.16).



**Abbildung 2.4**

Links: Wertepaare mit  $u_i < u_{\text{ref}}$ , wobei die Teilnehmerwerte gegeben sind durch  $x_i = 0$  und  $U_i = 1,0$ . Der Referenzwert wird variiert von  $x_{\text{ref}} = 0$  bis  $x_{\text{ref}} = -3$ . Die Referenzunsicherheit bleibt konstant mit  $U_{\text{ref}} = 2,0$ .

Rechts: Vergleich des  $E_{n;i}$ -Wertes aus Gleichung (2.8) mit dem modifizierten  $E_{n;i}^*$ -Wert aus Gleichung (2.16) unter Verwendung der Daten der linken Graphik. Im Bereich  $E_{n;i}^* \geq 1$  und  $E_{n;i} < 1$  profitiert der Teilnehmer bei einer Leistungsbewertung gemäß Gleichung (2.8) von der größeren Referenzunsicherheit  $u_{\text{ref}} > u_i$ .

### 2.2.3 Der $E_n$ -Wert als Hypothesentest

Das in Abschnitt 2.2 definierte Kriterium zur Leistungsbewertung eines Teilnehmers folgt dem statistischen Konzept eines Hypothesentests, genauer gesagt dem des zweiseitigen Signifikanztests, der sich mit der Differenz zweier Erwartungswerte befasst [14]. Um dieses Konzept nachfolgend anzuwenden, sei definiert, dass geschwungene Buchstaben in diesem Unterkapitel eine statistische Größe bezeichnen, wohingegen die Standard-Schriftart für konkrete Realisierungen der statistischen Größen verwendet wird. Beispielsweise bezeichnet  $\mathcal{E}_{n;i}$  eine statistische Größe, wohingegen  $E_{n;i}$  eine konkrete Realisierung, d.h. eine Stichprobe derselbigen darstellt. Letzteres wird bereits aus der Definition der Größe  $E_{n;i}$  in Gleichung (2.8) deutlich, in der die Erwartungswerte des Teilnehmer- und des Referenzergebnisses im Zähler des Bruches stehen.

Als Nullhypothese  $\mathbb{H}_0$  wird definiert

$$\mathbb{H}_0 : E(\mathcal{E}_{n;i}) = 0, \quad (2.18)$$

Hierbei ist  $E(\mathcal{E}_{n;i})$  als der Erwartungswert der statistischen Variable  $\mathcal{E}_{n;i}$  zu erachten, welche in Anlehnung an Gleichung (2.8) definiert ist durch

$$\mathcal{E}_{n;i} = \frac{\mathcal{D}_i}{2u(\mathcal{D}_i)}. \quad (2.19)$$

In Gleichung (2.19) ist  $\mathcal{D}_i$  als die statistische Variable der Differenz zwischen Teilnehmer- und Referenzwert zu verstehen, dessen konkrete Realisierung durch  $d_i$  aus Gleichung (2.6) gegeben ist. Anhand Gleichung (2.19) zeigt sich, dass  $\mathcal{E}_{n;i}$  durch die gleiche Wahrscheinlichkeitsdichte beschrieben werden kann wie  $\mathcal{D}_i$ , der Differenz zwischen Teilnehmerwert und Referenzwert. Da im Rahmen dieses Dokumentes stets von einer Normalverteilung ausgegangen wird, pflanzt sich diese Annahme auch in  $\mathcal{E}_{n;i}$  fort. Aufgrund der Normierung in Gleichung (2.19) gilt für die Standardabweichung von  $\mathcal{E}_{n;i}$  allgemein

$$u(\mathcal{E}_{n;i}) = \frac{u(\mathcal{D}_i)}{2u(\mathcal{D}_i)} = \frac{1}{2}. \quad (2.20)$$

Die zu Gleichung (2.18) gehörende Alternativhypothese  $\mathbb{H}_1$  lautet

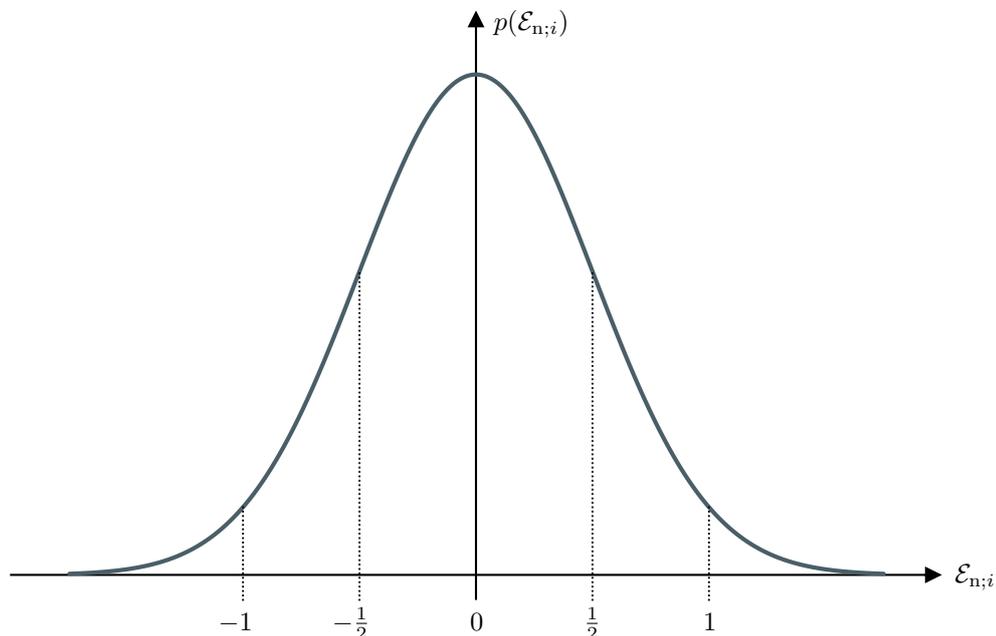
$$\mathbb{H}_1 : E(\mathcal{E}_{n;i}) \neq 0. \quad (2.21)$$

Der Hypothesentest wird zweiseitig durchgeführt, da positive oder negative Abweichungen zwischen Teilnehmer- und Referenzwert gleichwertig sind. Die Nullhypothese  $\mathbb{H}_0$  aus Gleichung (2.18) nimmt an, dass Teilnehmer- und Referenzwert übereinstimmen. Diese Annahme gilt es zu bestätigen oder zu verwerfen.

Unter der Annahme der Nullhypothese aus Gleichung (2.18), sowie der allgemeinen Standardabweichung aus Gleichung (2.20), als auch der Annahme einer Normalverteilung gilt für die statistische Größe  $\mathcal{E}_{n;i}$

$$p(\mathcal{E}_{n;i}) = \sqrt{\frac{2}{\pi}} e^{-2\mathcal{E}_{n;i}^2}. \quad (2.22)$$

Der konkrete  $\mathcal{E}_{n;i}$ -Wert, den ein Teilnehmer  $i$  in der Auswertung der Eignungsprüfung nach Gleichung (2.8) erzielt, wird als Stichprobe aus der in Abbildung 2.5 dargestellten Normalverteilung erachtet.



**Abbildung 2.5**

Für die statistische Größe  $\mathcal{E}_{n;i}$  wird eine Normalverteilung angenommen, mit normierter Standardabweichung  $1/2$ . Für den Erwartungswert wird Null angesetzt, entsprechend der Nullhypothese  $\mathbb{H}_0$ .

Das Signifikanzniveau des Hypothesentests wird gemäß [10] auf

$$\alpha = 4,55 \% \quad (2.23)$$

gesetzt. Da es sich um einen zweiseitigen Hypothesentest handelt, bedeutet dies, dass die Nullhypothese akzeptiert wird, wenn die  $E_{n;i}$ -Stichprobe des Teilnehmers  $i$  im Bereich  $1 - \alpha = 95,45 \%$  symmetrisch um den Erwartungswert der Normalverteilung aus Gleichung (2.22) liegt. Über das gewählte Signifikanzniveau kann ein kritischer Wert definiert werden, gemäß

$$\int_{-E_{n;krit}}^{E_{n;krit}} p(\mathcal{E}_{n;i}) d\mathcal{E}_{n;i} = \text{erf}(\sqrt{2}E_{n;krit}) = 1 - \alpha. \quad (2.24)$$

Unter Verwendung der Gaußschen Fehlerfunktion  $\text{erf}(z)$  mit

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-x^2} dx. \quad (2.25)$$

In Gleichung (2.24) wurde berücksichtigt, dass es sich um einen zweiseitigen Hypothesentest handelt. Für das in Gleichung (2.23) gewählte Signifikanzniveau resultiert

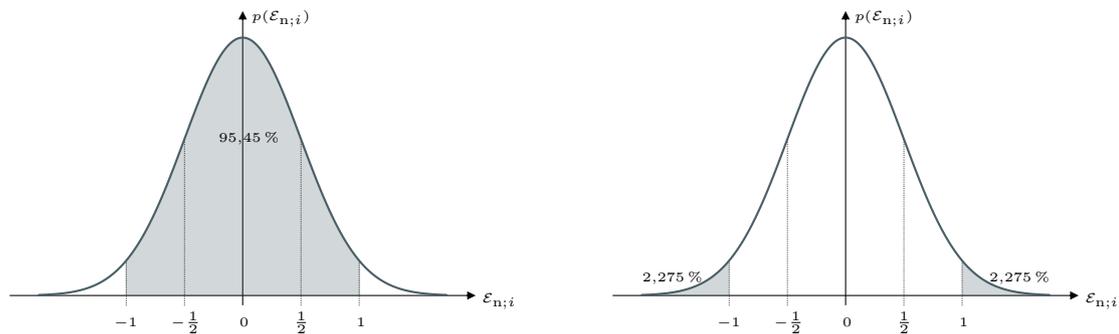
$$E_{n;krit} = 1. \quad (2.26)$$

Um die Nullhypothese zu bestätigen, d.h. letztlich die Annahme, dass Teilnehmer- und Referenzergebnis übereinstimmen und die Leistung des Teilnehmers  $i$  positiv bewertet werden kann, muss entsprechend der Zweiseitigkeit des Hypothesentests gelten

$$-E_{n;krit} \leq E_{n;i} \leq E_{n;krit} \quad (2.27)$$

Dies ist gleichbedeutend mit dem in Gleichung (2.9) definierten Leistungskriterium. Der Annahmehereich der Nullhypothese wird demnach aufgespannt durch das Intervall  $[-E_{n;krit}; E_{n;krit}]$ . Der Ablehnungsbereich ist entsprechend durch den restlichen Bereich gegeben, d.h. durch  $(-\infty; \infty) \setminus [-E_{n;krit}; E_{n;krit}]$ . Abbildung 2.6 zeigt die graphische Veranschaulichung des Annahme- und Ablehnungsbereiches der Nullhypothese.

In der definierten Nullhypothese aus Gleichung (2.18) steckt die Annahme, dass systematische Fehler im Teilnehmer- und Referenzwert korrigiert werden. Jegliche Abweichungen, die sich zwischen Teilnehmer- und Referenzwert ergeben sind dann als zufällig zu erachten, weshalb die konkrete Stichprobe  $E_{n;i}$  von der Erwartung  $E(\mathcal{E}_{n;i}) = 0$  abweichen darf. Diese Abweichung darf maximal  $E_{n;krit}$  betragen, was den Annahmehereich definiert. Hinter der Wahl der Annahme- und Ablehnungsbereiche steht der Gedanke, dass eine Stichprobe  $E_{n;i}$ , die betragsmäßig größer ist als der kritische Wert  $E_{n;krit}$  unter der Nullhypothese und der um Null zentrierten Normalverteilung eher unwahrscheinlich ist, obgleich diese Werte statistisch möglich sind. Das Signifikanzniveau definiert daher das Risiko, dass die Nullhypothese wahr ist, aber verworfen wird, aufgrund einer Stichprobe die im Bereich  $(-\infty; \infty) \setminus [-E_{n;krit}; E_{n;krit}]$  liegt. Je kleiner das Signifikanzniveau gewählt wird, desto kleiner wird das Risiko einer falsch-negativen Leistungsbewertung, d.h. einer fälschlichen Zurückweisung der Nullhypothese. Im Falle von Eignungs-



**Abbildung 2.6**

Links: Gezeigt ist der Annahmereich der Nullhypothese. Der kritische Wert ist entsprechend Gleichung (2.24) durch  $E_{n;krit} = 1$  gegeben und liegt zweiseitig, symmetrisch um den Erwartungswert  $E(\varepsilon_{n;i}) = 0$  aus der Nullhypothese  $\mathbb{H}_0$ . Die Wahrscheinlichkeit des Annahmereichs ist gegeben durch  $1 - \alpha = 95,45\%$ .

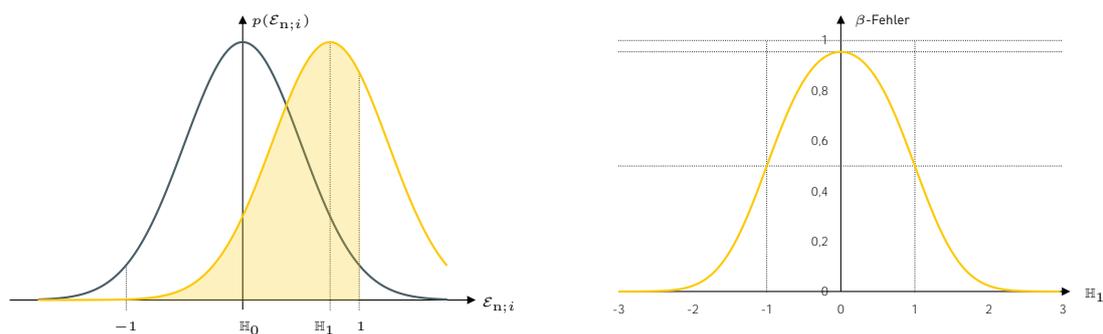
Rechts: Die grau hinterlegte Fläche zeigt den Ablehnungsbereich.  $E_{n;i}$ -Realisierungen eines Teilnehmers, die in diesen Bereich fallen führen zu einer Verwerfung der Nullhypothese. Der Bereich ist links und rechts durch  $\alpha/2 = 2,275\%$  gegeben.

prüfungen wird gemäß [10] ein Signifikanzniveau, d.h. ein maximales falsch-negativ Risiko von 4,55% gewählt. Das falsch-negativ Risiko wird im Rahmen von Hypothesentests auch als  $\alpha$ -Fehler oder Fehler 1. Art bezeichnet [15]. Dem gegenüber steht der sogenannte  $\beta$ -Fehler oder Fehler 2. Art, der das falsch-positiv Risiko bezeichnet, d.h. die fälschliche Bestätigung der Nullhypothese [15]. Die nachfolgende Tabelle zeigt die möglichen Entscheidungen anhand der Stichprobe  $E_{n;i}$  und deren Bewertung vor dem Hintergrund der tatsächlichen Eigenschaft der Grundgesamtheit.

Stichprobe	Grundgesamtheit	
	$\mathbb{H}_0$ ist wahr	$\mathbb{H}_0$ ist falsch ( $\mathbb{H}_1$ ist wahr)
$-1 < E_{n;i} < 1$  $\mathbb{H}_0$ wird angenommen	korrekte positive Leistungsbewertung	falsch-positiv Leistungsbewertung  $\beta$ -Fehler (i.A. nicht näher angebar)
$-1 \geq E_{n;i} \geq 1$  $\mathbb{H}_0$ wird abgelehnt ( $\mathbb{H}_1$ wird angenommen)	falsch-negativ Leistungsbewertung  $\alpha$ -Fehler (vgl. Signifikanzniveau $\alpha$ )	korrekte negative Leistungsbewertung

Da die Alternativhypothese  $\mathbb{H}_1$  im Fall des zweiseitigen Signifikanztests nicht konkretisiert wird, sondern lediglich die Negierung der Nullhypothese ist, kann das falsch-positiv Risiko i.A. nicht berechnet werden. Letztlich kann das Risiko der falschen Annahme der Nullhypothese sehr groß werden, wie

Abbildung 2.7 zeigt. Zur Veranschaulichung des  $\beta$ -Fehlers, wurde in Abbildung 2.7 angenommen, dass eine konkrete Alternativhypothese  $\mathbb{H}_1$  vorliegt, bspw.  $E(\mathcal{E}_{n;i}) = 0,75$ . Aufgrund der Normierung ist die Standardabweichung der Alternativhypothese weiterhin durch  $1/2$  gegeben, sodass die Alternativhypothese  $\mathbb{H}_1$  eine zur Nullhypothese verschobene Normalverteilung ist. Der  $\beta$ -Fehler ist dann die Fläche der Alternativhypothese, im Bereich  $[-E_{n;krit}; E_{n;krit}]$ . Dadurch wird das Risiko berechnet, dass die Nullhypothese angenommen wird (d.h.  $-E_{n;krit} \leq E_{n;i} \leq E_{n;krit}$ ), obgleich die Alternativhypothese wahr ist. Wie Abbildung 2.7 zeigt, wird dieser Fehler gerade dann sehr groß, wenn Null- und Alternativhypothese sehr nahe beieinander sind. Da im vorliegenden Fall die Alternativhypothese nicht weiter konkretisiert ist, kann der  $\beta$ -Fehler nicht berechnet werden. Die vorhergehenden Ausführungen machen jedoch deutlich, dass dieser prinzipiell zwischen 0 % und 94,45 % liegen kann. Die Annahme der Nullhypothese ist demnach typischerweise mit einem größeren Risiko verbunden als die Ablehnung der Nullhypothese. Der Hypothesentest ist daher so formuliert, dass negative Leistungsbewertungen eines Teilnehmers  $i$  ein geringes Risiko aufweisen und durch das gewählte Signifikanzniveau begrenzt sind. Je kleiner das gewählte Signifikanzniveau ist, desto kleiner wird das falsch-negativ Risiko. Die Intention des zweiseitigen Signifikanztests ist daher nicht die Nullhypothese zu bestätigen, sondern sie zu verwerfen.



**Abbildung 2.7**

*Links: Unter der Annahme, dass eine konkrete Alternativhypothese (gelb) vorliegt, kann der  $\beta$ -Fehler, d.h. das Risiko einer fälschlichen Annahme der Nullhypothese (grau) bestimmt werden und ist durch die Fläche der Alternativhypothese gegeben, die zwischen  $\pm E_{n;krit} = \pm 1$  liegt. Im Beispiel wurde als Alternativhypothese  $E(\mathcal{E}_{n;i}) = 0,75$  gewählt. Siehe auch [15].*

*Rechts: Gezeigt ist der  $\beta$ -Fehler für verschiedene Alternativhypothesen  $\mathbb{H}_1$ . Je ähnlicher sich Nullhypothese und Alternativhypothese sind, desto größer wird das Risiko der fälschen Annahme der Nullhypothese.*

Mit den Ausführungen dieses Unterkapitels, lauten die zwei Aussagen, die sich durch die Leistungsbewertung in einer Eignungsprüfung anhand des zweiseitigen Signifikanztests treffen lassen

1.  $-1 \leq E_{n;i} \leq 1$ :

Die Nullhypothese kann nicht verworfen werden, da die Stichprobe einen Wert aus dem 95,45 %-igen Annahmebereich um die Nullhypothese zeigt. Die Leistung des Teilnehmers wird positiv bewertet. Das Risiko dieser Entscheidung kann nicht näher bestimmt werden.

2.  $|E_{n;i}| > 1$ :

Die Nullhypothese wird verworfen, da die Stichprobe einen Wert aus dem 4,55 %-igen Ablehnungs-

bereich zeigt. Die Leistung des Teilnehmers wird negativ bewertet. Das Risiko dieser Entscheidung liegt bei maximal  $\alpha = 4,55\%$ .

Anmerkung 1: Neben der Prüfung der Stichprobe  $E_{n;i}$  auf einen kritischen Wert  $E_{n;krit}$ , kann im Rahmen von Hypothesentests auch der  $p$ -Wert berechnet werden. Dieser ist für einen zweiseitigen Test definiert durch

$$\begin{aligned}
 p &= 1 - 2 \int_{-E_{n;i}}^{E_{n;i}} p(\mathcal{E}_{n;i}) d\mathcal{E}_{n;i} \\
 &= 1 - \operatorname{erf}(\sqrt{2}E_{n;i})
 \end{aligned}
 \tag{2.28}$$

Unter Verwendung der Gaußschen Fehlerfunktion aus Gleichung (2.25). Die Nullhypothese wird verworfen, wenn  $p < \alpha$  ist und angenommen, wenn  $p \geq \alpha$ .

Anmerkung 2: Der Hypothesentest des  $E_{n;i}$ -Wertes ist in Anlehnung an den Gauß-Test [16], auch  $z$ -Test genannt, für eine einzige Stichprobe definiert. Der Unterschied liegt darin, dass die Differenz  $d_i$  nicht auf die einfache Standardabweichung, sondern auf die mit  $k = 2$  erweiterte Messunsicherheit normiert wird. Dies führt lediglich zu einer Reskalierung, sodass der kritische Wert für das Signifikanzniveau aus Gleichung (2.23) durch  $E_{n;krit} = 1$  gegeben ist, wohingegen der entsprechende kritische  $z$ -Wert  $z_{krit} = 2$  beträgt.

#### 2.2.4 Berücksichtigung von Korrelationen im $E_n$ -Wert

Die obigen Ausführungen haben gezeigt, dass die Leistungsbewertung eines Teilnehmers durch das in Gleichung (2.8) und Gleichung (2.9) definierte Leistungskriterium, die Kenntnis eines Referenzwertes  $x_{ref}$  und einer Referenzunsicherheit  $U_{ref} = 2u_{ref}$  voraussetzt. Ein zentraler Bestandteil der statistischen Auswertung einer Eignungsprüfung, und somit dieses Dokuments, ist folglich die Bestimmung dieser Referenzdaten. In den nachfolgenden Kapiteln werden vier Methoden zur Bestimmung der Referenzdaten vorgestellt. Diese sind sukzessive bezeichnet als Methode A (Kapitel 3), Methode B (Kapitel 4), Methode C (Kapitel 5) und Methode D (Kapitel 6). Methode A bezeichnet die weit verbreitete Vorgehensweise, den Referenzwert über die Messungen eines Referenzlabors bestimmen zu lassen. Dem gegenüber stehen Methode B, C und D. Hier werden statistische Verfahren genutzt, um die Referenzdaten aus den Teilnehmerdaten zu berechnen. In diesen Fällen ist die in Gleichung (2.7) enthaltene Korrelation zwischen Teilnehmerdaten und Referenzdaten zu berücksichtigen.

Betrachtet werden im vorliegenden Dokument Modelle, in denen ein linearer Zusammenhang zwischen Teilnehmerdaten und Referenzdaten vorliegt, d.h.

$$x_{ref} = \sum_{i=1}^N w_i x_i.
 \tag{2.29}$$

Für die Referenzunsicherheit  $u_{ref}$  gilt im Modell aus Gleichung (2.29)

$$u_{\text{ref}} = \sqrt{\sum_{i=1}^N w_i^2 u_i^2}. \quad (2.30)$$

Die Faktoren  $w_i$  in Gleichung (2.29) können als Gewicht bezeichnet werden, da dadurch der Anteil des Teilnehmerwertes  $x_i$ , der in den Referenzwert einfließt festgelegt ist. Die Summe aller Gewichte muss normiert sein, d.h.

$$\sum_{i=1}^N w_i = 1. \quad (2.31)$$

Zudem gilt

$$0 \leq w_i \leq 1. \quad (2.32)$$

Der Grenzfall  $w_i = 1$  entspricht Methode A, d.h. dem Fall, dass die Messung des Labors  $i$  die Referenzdaten definiert. Gilt für einen Teilnehmer  $w_i = 0$ , so fließen dessen Daten nicht in die Bestimmung der Referenzdaten ein. Für die Entscheidung, ob ein Teilnehmer in die Bestimmung der Referenzdaten einfließt oder nicht, gibt es üblicherweise definierte Kriterien, die sich jedoch je nach Auswertungsmethode unterscheiden können. Die Gesamtzahl der Teilnehmer kann prinzipiell in zwei Gruppen aufgeteilt werden. Gruppe „r“ enthält die  $N_r$  Teilnehmer, deren Daten in der Berechnung der Referenzdaten berücksichtigt wird. Gruppe „n“ enthält die  $N_n = N - N_r$  Teilnehmer, die in der Berechnung nicht berücksichtigt werden. Der Gewichtungsfaktor eines Teilnehmers  $i$  aus Gruppe „n“ ist durch  $w_i = 0$  gegeben.

Im linearen Modell von Gleichung (2.29) und unter der Verwendung der Normierungsbedingung aus Gleichung (2.31) kann leicht gezeigt werden, dass

$$\min_{i \in \text{Gruppe r}} (x_i) \leq x_{\text{ref}} \leq \max_{i \in \text{Gruppe r}} (x_i), \quad (2.33)$$

d.h. der resultierende Referenzwert liegt immer zwischen dem minimalen und dem maximalen Teilnehmerwert aus „r“. Für die Referenzunsicherheit gilt folgende Abschätzung

$$\sqrt{N_r} \cdot \min_{i \in \text{Gruppe r}} (u_i) \cdot \min_{i \in \text{Gruppe r}} (w_i) \leq u_{\text{ref}} \leq \sqrt{N_r} \cdot \max_{i \in \text{Gruppe r}} (u_i) \cdot \max_{i \in \text{Gruppe r}} (w_i). \quad (2.34)$$

Für das in Gleichung (2.29) angegebene Modell, ist die Kovarianz zwischen Teilnehmerwert  $x_j$  und Referenzwert gegeben durch

$$u(x_i; x_{\text{ref}}) = w_i \cdot u_i^2. \quad (2.35)$$

Damit lässt sich Gleichung (2.7) ausdrücken als

$$\begin{aligned}
 U_{d_i} &= 2\sqrt{(1-2w_i)u_i^2 + u_{\text{ref}}^2} \\
 &= 2\sqrt{(1-w_i)^2u_i^2 + \sum_{\substack{j=1 \\ j \neq i}}^N w_j^2 u_j^2}
 \end{aligned} \tag{2.36}$$

Hierbei wurde in der zweiten Zeile die Definition der Referenzunsicherheit aus Gleichung (2.30) genutzt. Es zeigt sich insbesondere, dass die Unsicherheit  $U_{d_i}$  immer definiert ist, da der Ausdruck unter der Wurzel nicht negativ werden kann. Zu beachten ist jedoch, dass  $U_{d_i} = 0$ , falls  $w_i = 1$  und  $w_j = 0 \quad \forall j \neq i \in [1; N]$ , wobei letzteres ein Resultat der Normierung aus Gleichung (2.31) ist. Der Grenzfall  $U_{d_i} = 0$  ist problematisch, da die Unsicherheit aus Gleichung (2.36) im Nenner des  $E_{n;i}$ -Wertes steht. In diesem Fall wird jedoch auch der Zähler des  $E_{n;i}$ -Wertes gleich Null, da  $x_{\text{ref}} = x_i$ . Letztlich spiegelt dies somit gerade den Fall wieder, dass die Messungen des Teilnehmers  $i$  die Referenzdaten definieren, d.h. Auswertungsmethode A verwendet wird, die in Kapitel 3 beschrieben ist. Für Teilnehmer  $i$  ist eine  $E_{n;i}$ -Auswertung daher nicht sinnvoll und auch nicht vorgesehen. In allen anderen Fällen ist der  $E_{n;i}$ -Wert als Variable zur Leistungsbewertung eindeutig definiert und gegeben durch

$$E_{n;i} = \frac{x_i - x_{\text{ref}}}{2\sqrt{(1-2w_i)u_i^2 + u_{\text{ref}}^2}}. \tag{2.37}$$

Ergänzend zu Gleichung (2.35) ist nachfolgend noch der Korrelationskoeffizient  $r_{i;\text{ref}}$  angegeben

$$r_{i;\text{ref}} = \frac{u(x_i; x_{\text{ref}})}{u_i u_{\text{ref}}} = w_i \cdot \frac{u_i}{u_{\text{ref}}}. \tag{2.38}$$

Korrelationskoeffizienten werden im allgemeinen Werte zwischen 0 und 1 zugeordnet. Da alle Parameter in Gleichung (2.38) per Definition größer oder gleich null sind, ist zunächst klar, dass auch  $r_{i;\text{ref}} \geq 0$ . Zu zeigen ist noch, dass  $r_{i;\text{ref}} \leq 1$ . Hierfür kann verwendet werden, dass  $r_{i;\text{ref}} \leq 1 \leftrightarrow r_{i;\text{ref}}^2 \leq 1$ . Daraus folgt

$$\begin{aligned}
 r_{i;\text{ref}}^2 &= \frac{w_i^2 u_i^2}{u_{\text{ref}}^2} \\
 &= \frac{w_i^2 u_i^2}{w_i^2 u_i^2 + \sum_{\substack{j=1 \\ j \neq i}}^{N_r} w_j^2 u_j^2} \\
 &\leq \frac{w_i^2 u_i^2}{w_i^2 u_i^2} = 1.
 \end{aligned} \tag{2.39}$$

Damit gilt für alle linearen Modelle, dass der Korrelationskoeffizient wohl definiert ist. In den nachfolgenden Kapiteln werden verschiedenen Methoden zur Bestimmung der Referenzdaten diskutiert. Die Charakteristik der jeweiligen Methode hängt dabei insbesondere von der Definition der Gewichtungsfaktoren  $w_i$  ab. Diese haben sowohl einen Einfluss auf den Referenzwert und dessen Unsicherheit, als auch auf die Leistungsbewertung über den  $E_{n;i}$ -Wert. Darauf wird in den nun folgenden Kapiteln eingegangen.

Anmerkung 1: Im Rahmen dieses Dokuments wird für Methode B, C und D die Bezeichnung verwendet, dass Teilnehmerdaten in der Bestimmung der Referenzdaten berücksichtigt werden. An dieser Stelle sei explizit darauf hingewiesen, dass die Methoden B, C und D so zu verstehen sind, dass die Referenzdaten aus den Messergebnissen mehrere Labore berechnet werden, wohingegen bei Methode A die Referenzdaten durch die Messungen eines Labors festgelegt sind. Zudem müssen bei Methode B, C und D nicht zwangsläufig alle Teilnehmerdaten in die Bestimmung der Referenzdaten einfließen. Vielmehr kann auch eine Vorauswahl anhand zusätzlicher Kriterien getroffen werden. Hierauf wird am Ende von Kapitel 3 eingegangen. Aus Gründen der Lesbarkeit wird der Begriff Teilnehmer für alle Labore verwendet, unabhängig davon, ob sie letztlich Einfluss auf die Referenzdaten nehmen oder nicht. Letzteres wird durch die Zuordnung zu Gruppe „r“ oder „n“ berücksichtigt.

Anmerkung 2: Bei der Bestimmung von Referenzdaten anhand von Teilnehmerdaten muss die Rückführung der Referenzdaten gemäß DIN EN ISO/IEC 17043:2023 [2] in einer Eignungsprüfung gewährleistet sein. Aus diesem Grund können nur Teilnehmer, die in der Eignungsprüfung am entsprechenden Messpunkt eine akkreditierte Messunsicherheit berichten in die Bestimmung der Referenzdaten einfließen. Berichtet ein akkreditierter Teilnehmer eine Messunsicherheit, die kleiner ist als seine akkreditierte Messunsicherheit, so kann er nicht für die Bestimmung der Referenzdaten herangezogen werden. Insbesondere ist es nicht möglich bei einem solchen Teilnehmer die akkreditierte Messunsicherheit zur Bestimmung der Referenzdaten zu verwenden um damit dann die Leistung anhand der berichteten, nicht-akkreditierten Messunsicherheit zu bewerten. Der Grund dafür ist, dass diese zu nicht-definierten  $E_n$ -Werten in Gleichung (2.37) führen kann, aufgrund der Korrelation. Deutlich wird dies vor allem in Gleichung (4.11). Liegt eine Eignungsprüfung mit nur nicht-akkreditierten Teilnehmern vor, so empfiehlt sich die Verwendung von Methode A. Zusätzlich sei an dieser Stelle darauf hingewiesen, dass bei der Berechnung der Referenzdaten durch Teilnehmerdaten darauf zu achten ist, ob Korrelationen zwischen den Teilnehmerdaten bestehen, beispielsweise durch eine Rückführung auf das gleiche Labor. Solche Korrelationen können vor allem bei kleinen Teilnehmerzahlen den Referenzwert beeinflussen und sollten daher vermieden werden.

Anmerkung 3: Die Berücksichtigung von Teilnehmerdaten in den Methoden B, C und D ersetzt nicht die, durch die DIN EN ISO/IEC 17043:2023 [2] vorgeschriebenen, Stabilitätsuntersuchungen durch den Eignungsprüfungsanbieter. Durch zusätzliche Stabilitätsmessungen vor, zwischen und/oder nach den Teilnehmersmessungen ist die Stabilität des Eignungsprüfungsgegenstandes vor und während der Durchführung der Eignungsprüfung zu untersuchen und auszuwerten. Hierfür sind geeignete Stabilitätskriterien heranzuziehen und ggf. entsprechende Stabilitätsbeiträge abzuleiten. Im Rahmen dieses Dokuments wird hierauf nicht weiter eingegangen.

### 3 Methode A — Referenzlabor

Gemäß DIN EN ISO/IEC 17043:2023 [2] und ISO 13528:2022 [3] ist eine Möglichkeit zur Bestimmung der Referenzdaten durch die Wahl eines Referenzlabors gegeben. In der Praxis ist dieses Vorgehen durch der leichten Anwendbarkeit populär und aufgrund des direkten Vergleiches zweier Messergebnisse auch allgemein anerkannt. Nach definierten Kriterien wird ein für die Messgröße  $X$  geeignetes Referenzlabor bestimmt, das dann durch eine Messung die Referenzdaten ermittelt. Die Kriterien für die Wahl des Referenzlabors können vielfältig sein und sich je nach Ziel der Eignungsprüfung unterscheiden. Nach DIN EN ISO/IEC 17043:2023 [2], Absatz A.2 kann ein solches Kriterium lauten, dass „ein Referenzlaboratorium [...] eingesetzt [wird], das in der Lage ist, für den [...] [Eignungsprüfungs-]gegenstand metrologisch rückgeführte [...] Werte mit hinreichend geringer Messunsicherheit und Zuverlässigkeit zu liefern.“. Insbesondere sind hierfür Nationalinstitute, aber auch nach DIN EN ISO/IEC 17025:2018 [1] akkreditierte Kalibrierlabore geeignet. Je nach Messgröße und Anforderung an die Eignungsprüfung können auch weitere Kriterien für die Wahl des Referenzlabors herangezogen werden. Im Rahmen dieses Dokuments wird darauf jedoch nicht weiter eingegangen.

In der vorliegende Methode A wird auf den in Abschnitt 2.1 beschriebenen Datensatz, bestehend aus  $N$  Teilnehmerdaten, zurück gegriffen. Ohne Beschränkung der Allgemeinheit sei der Datensatz des Referenzlabors durch den Index  $i = N$  bezeichnet und besteht aus

- dem Kalibrierwert  $x_{N;0} \equiv x_{\text{ref};0}$  (s. dazu auch die Anmerkung unter 2.1),
- dem Messwert  $x_{N;\text{meas}} \equiv x_{\text{ref};\text{meas}}$
- und der dazugehörigen erweiterten Messunsicherheit  $U_N \equiv U_{\text{ref}}$

des gewählten Referenzlabors und definiert damit die benötigten Referenzdaten für die Leistungsbewertung der verbleibenden  $N - 1$  Teilnehmer. Entsprechend Abschnitt 2.1 und der Definition des Teilnehmerdatensatzes ist auch hier die Messabweichung  $x_{\text{ref}}$  mit

$$x_{\text{ref}} = x_{\text{ref};\text{meas}} - x_{\text{ref};0} \quad (3.1)$$

die benötigte Größe für die Auswertung der Eignungsprüfung. Unter der Annahme einer Normalverteilung, liegt der erweiterten Messunsicherheit des Referenzlabors typischerweise ein Erweiterungsfaktor  $k_{\text{ref}} = 2$ , d.h. eine Überdeckungswahrscheinlichkeit von ca. 95,45 %, zugrunde, sodass

$$U_{\text{ref}} = 2 \cdot u_{\text{ref}}. \quad (3.2)$$

Mit  $u_{\text{ref}}$  ist die Standardunsicherheit des Referenzlabors bezeichnet. Für die in Gleichung (2.7) berücksichtigte Kovarianz wird typischerweise die Annahme

$$u(x_i; x_{\text{ref}}) = 0 \quad \forall i \in [1; N - 1] \quad (3.3)$$

getroffen, d.h. es wird für alle verbleibenden  $N - 1$  Teilnehmerwerte  $x_i$  angenommen, dass diese mit dem Referenzwert  $x_{\text{ref}}$  nicht korreliert sind. Im Modell von Abschnitt 2.2.4 entspricht dies dem Grenzfall

$w_N = 1$  und  $w_i = 0 \forall i \in [1; N - 1]$ . Diese Annahme führt dazu, dass sich der  $E_{n;i}$ -Wert aus Gleichung (2.8) vereinfacht ausdrücken lässt durch

$$E_{n;i} = \frac{x_i - x_{\text{ref}}}{2\sqrt{u_i^2 + u_{\text{ref}}^2}} = \frac{x_i - x_{\text{ref}}}{\sqrt{U_i^2 + U_{\text{ref}}^2}}. \quad (3.4)$$

Das Kriterium zur Leistungsbewertung aus Gleichung (2.9) bleibt bestehen. Wie in Abschnitt 2.2.4 beschrieben, ist eine Leistungsbewertung für das gewählte Referenzlabor nicht möglich. Die Annahme, dass zwischen Teilnehmerwert und Referenzwert keine Korrelation vorliegt, ist insbesondere dann zu überprüfen, wenn die Rückführung eines Teilnehmers  $i$  für die Messgröße  $X$  über das gewählte Referenzlabor erfolgt. Dies ist entweder in der Planung und Organisation der Eignungsprüfung zu berücksichtigen, oder durch die Kovarianz im  $E_{n;i}$ -Wert gemäß Gleichung (2.8). Unter Umständen ist die Berechnung dieser Kovarianz nicht trivial, da detaillierte Kenntnisse über die Bestimmung der Teilnehmer- und Referenzwerte, sowie deren Unsicherheiten bekannt sein müssen.

Im eingangs zitierten Abschnitt A.2 aus DIN EN ISO/IEC 17043:2023 [2], impliziert „mit hinreichend geringer Messunsicherheit“, dass bei der Wahl des Referenzlabors darauf zu achten ist, dass die durch das Referenzlabor gestellte Referenzunsicherheit idealerweise kleiner ist als die Messunsicherheit des vorliegenden bzw. zu erwartenden Teilnehmerkreises. Leistungsbewertungen für Teilnehmer, deren Messunsicherheit kleiner ist als die Referenzunsicherheit sind mathematisch nach Gleichung (2.8) und Gleichung (2.9) möglich, jedoch ist die Aussagekraft dieses Ergebnisses eingeschränkt. In einem solchen Fall kann eine Auswertung nach Methode B (Kapitel 4), Methode C (Kapitel 5) oder Methode D (Kapitel 6) eine geeignete Möglichkeit darstellen um diese Unzulänglichkeit zu beheben.

Aus mathematisch-statistischer Sicht kann Methode A so zusammengefasst werden, dass die Referenzdaten definiert werden und durch die Wahl des Referenzlabors und dessen Messungen festgelegt sind. Dem gegenüber stehen statistische Auswertungsmethoden, die in den nachfolgenden Kapiteln diskutiert werden. An dieser Stelle sei angemerkt, dass auch die Kombination von Methode A mit einer statistischen Methode (bspw. die in diesem Dokument vorgestellten Methoden B, C oder D) möglich ist. In diesem Fall wird der Referenzwert aus den Messergebnissen von  $N_r$  ausgewählten Teilnehmern festgelegt. Die Messungen dieser  $N_r$  Teilnehmer sind in einer geeigneten Weise im Eignungsprüfungsablauf unter zu bringen. Die genannte Vorgehensweise entspricht im Wesentlichen der Anwendung einer statistischen Methode, unter der zusätzlichen Definition eines qualitativen Ausschlusskriteriums, sodass nur die Daten der gewählten  $N_r$  Teilnehmer in die Berechnung einfließen. Für die restlichen  $N - N_r$  Teilnehmer ändert sich im Vergleich zu Methode A nichts, da sie keinen Einfluss auf die Referenzdaten haben. Durch die statistische Auswertung der  $N_r$  Teilnehmer wird die Bestimmung und Kenntnis der Referenzdaten jedoch verbessert. Je nach Datenlage und gewählter statistischer Methode können möglicherweise auch kleinere Referenzunsicherheiten erzielt werden. Gewissermaßen entspricht die statistische Auswertung von  $N_r$  Teilnehmer einer „Eignungsprüfung in der Eignungsprüfung“, in dem Sinne, dass die  $N_r$  Teilnehmer sich ebenfalls mit dem statistisch ermittelten Referenzwert vergleichen können. Bei einer reinen Auswertung nach Methode A, ist eine Leistungsbewertung für das gewählte Referenzlabor nicht möglich.

In den nachfolgenden Kapiteln wird der Fokus auf die statistische Bestimmung der Referenzdaten unter

der Berücksichtigung von Teilnehmerdaten gelegt. Für den Fall einer bilateralen Eignungsprüfung wird in Kapitel 8 gezeigt, dass eine sinnvolle Auswertung nur durch Methode A gegeben ist und die statistischen Methoden nicht anwendbar sind.

## 4 Methode B — Gewichteter Mittelwert (Cox)

Die in diesem Kapitel vorgestellte und als Methode B bezeichnete statistische Auswertung von Eignungsprüfungsdaten stammt aus der Veröffentlichung „Evaluation of key comparison data“ von Maurice Cox [4] und ist dort unter „*Procedure A*“ beschrieben. In Methode A, die in Kapitel 3 beschrieben ist, werden die für die Leistungsbewertung benötigten Referenzdaten aus mathematischer Sicht lediglich definiert. Im Gegensatz dazu steht Methode B, da die Referenzdaten in diesem Fall nicht durch die Messungen eines Labors festgelegt werden, sondern durch die Messungen vieler Labore und der anschließenden statistischen Verrechnung dieser Messdaten zu einem Referenzdatensatz. Somit werden der Referenzwert und dessen Unsicherheit in Methode B unter der Berücksichtigung von Teilnehmerdaten berechnet.

Methode B ist auch unter der Bezeichnung „gewichteter Mittelwert“ bekannt, da, wie nachfolgend dargestellt, die Messabweichung  $x_i$  eines Teilnehmers  $i$  entsprechend der dazugehörigen Teilnehmervarianz  $u_i^2$  gewichtet wird. Allerdings ist die Bezeichnung „gewichteter Mittelwert“ anwendbar auf eine Vielzahl von Berechnungsmethoden, und insbesondere für die in Abschnitt 2.2.4 unter Gleichung (2.29) definierten linearen Modelle allgemeingültig. Neben Methode B ist der Referenzwert auch in Methode C (Kapitel 5) und Methode D (Kapitel 6) berechnet als gewichteter Mittelwert. Charakteristisch für die jeweilige Methode ist wie die Gewichtungsfaktoren  $w_i$  aus Gleichung (2.29) konkret berechnet werden. Aus diesem Grund wird die in [4] beschriebene Methode in diesem Dokument nicht als „gewichteter Mittelwert“ bezeichnet, sondern als Methode B.

Bei Methode B handelt es sich um eine unsicherheitsbasierte Auswertungsmethode mit individueller Gewichtung. Diese Charakteristik resultiert daraus, dass der Gewichtungsfaktor  $w_i$  eines Teilnehmers  $i$  im wesentlichen durch dessen Standardunsicherheit  $u_i$  gegeben ist. Die Teilnehmerunsicherheiten sind in Methode B somit von zentraler Bedeutung, wie die nun folgenden Abschnitte veranschaulichen.

### 4.1 Berechnung der Referenzdaten

Betrachtet wird der in Abschnitt 2.1 definierte Datensatz bestehend aus den Messabweichungen  $x_i$  der Teilnehmer  $i \in [1; N]$ , sowie den dazugehörigen erweiterten Messunsicherheiten  $U_i$ . Wie bereits in Kapitel 2 beschrieben, wird im Rahmen dieses Dokuments angenommen, dass den einzelnen Teilnehmerergebnissen eine Normalverteilung zugrunde liegt und  $U_i = 2 \cdot u_i \quad \forall i \in [1; N]$ . Diese Annahme ist eine zentrale Voraussetzung für die Anwendung von Methode B.

Unter dieser Voraussetzung können prinzipiell alle  $N$  Teilnehmerdaten in die Bestimmung der Referenzdaten einfließen, solange der in 4.1.1 beschriebene  $\chi^2$ -Test bestanden ist. Wie in Kapitel 3 beschrieben ist es jedoch auch möglich, nur Teilnehmerdaten in die Bestimmung der Referenzdaten einfließen zu lassen, die gewisse „qualitative“ Kriterien erfüllen. Ein mögliches Kriterium kann beispielsweise sein, dass für die betrachtete Messgröße  $X$  nur nach DIN EN ISO/IEC 17025:2018 akkreditierte Labore berücksichtigt werden. Abweichend von der Gesamtzahl  $N$  der Teilnehmer kann somit die effektive Anzahl der Teilnehmer, die in die Bestimmung der Referenzdaten einfließt auch kleiner sein. Nachfolgend wird die Einteilung der Teilnehmer in Gruppe „r“ und Gruppe „n“ aus Abschnitt 2.2.4 verwendet. Gruppe „r“ enthält die  $N_r$  Teilnehmer, deren Daten in der Berechnung der Referenzdaten berücksichtigt wird.

Gruppe „n“ enthält die  $N_n = N - N_r$  Teilnehmer, die in der Berechnung nicht berücksichtigt werden. Der Gewichtungsfaktor eines Teilnehmers  $i$  aus Gruppe „n“ ist durch  $w_{i;B} = 0$  gegeben.

Wie eingangs bereits erwähnt zeichnet sich Methode B dadurch aus, dass der Wert  $x_i$  eines Teilnehmers  $i$  aus Gruppe „r“ entsprechend seiner Teilnehmervarianz  $u_i^2$  gewichtet wird. Die Gewichtungsfaktoren aus Gleichung (2.29) sind für Methode B folglich zunächst durch

$$w_{i;B} = \begin{cases} \frac{b}{u_i^2} & \text{Teilnehmer in Gruppe „r“,} \\ 0 & \text{Teilnehmer in Gruppe „n“} \end{cases} \quad (4.1)$$

definiert. Der Faktor  $b$  kann über die Normierungsbedingung aus Gleichung (2.31) bestimmt werden. Das Ergebnis lautet

$$b = \left( \sum_{i=1}^{N_r} \frac{1}{u_i^2} \right)^{-1}. \quad (4.2)$$

Betrachtet man nun Gleichung (2.30), so kann die Referenzunsicherheit im linearen Modell durch die Kenntnis der Gewichtungsfaktoren berechnet werden, es gilt:

$$\begin{aligned} u_{\text{ref};B}^2 &= \sum_{i=1}^N w_{i;B}^2 u_i^2 \\ &= \left( \sum_{i=1}^{N_r} \frac{1}{u_i^2} \right)^{-1} \end{aligned} \quad (4.3)$$

Damit folgt, dass  $b = u_{\text{ref};B}^2$ , sodass für den Gewichtungsfaktor  $w_{i;B}$  eines Teilnehmers  $i \in [1; N]$  gilt

$$w_{i;B} = \begin{cases} \frac{u_{\text{ref};B}^2}{u_i^2} & \text{Teilnehmer in Gruppe „r“,} \\ 0 & \text{Teilnehmer in Gruppe „n“.} \end{cases} \quad (4.4)$$

Damit ist letztlich auch die Berechnung des Referenzwertes gemäß Gleichung (2.29) gegeben durch

$$\begin{aligned} x_{\text{ref};B} &= \sum_{i=1}^N w_{i;B} x_i \\ &= u_{\text{ref};B}^2 \sum_{i=1}^{N_r} \frac{x_i}{u_i^2}. \end{aligned} \quad (4.5)$$

Gleichungen (4.3), (4.4) und (4.5) definieren gewissermaßen Methode B. Insbesondere wird durch die Definition der Gewichtungsfaktoren in Gleichung (4.4) klar, dass den Teilnehmerunsicherheiten  $u_i$  eine große Bedeutung zu kommt. Die Streuung der Teilnehmerwerte  $x_i$  wird in der Referenzunsicherheit nicht berücksichtigt. In diesem Sinne handelt es sich bei Methode B um eine unsicherheitsbasierte Auswertung. Mit der Verwendung der Teilnehmerunsicherheiten in den Gewichtungsfaktoren, liegt zudem eine individuelle Gewichtung vor. So wird beispielsweise ein Teilnehmer, dessen Unsicherheit

im Vergleich zu den anderen Teilnehmerunsicherheiten klein ist, eine starke Gewichtung erhalten. Hingegen werden Teilnehmer mit verhältnismäßig großen Unsicherheiten einen geringeren Einfluss auf den Referenzwert haben. Diese Charakteristik sowie weitere Eigenschaften von Methode B sind in Abschnitt 4.2 dargestellt.

#### 4.1.1 $\chi^2$ -Test

Wie bereits zu Beginn des vorherigen Abschnitts angemerkt, ist eine Voraussetzung der Verwendung von Methode B das Bestehen des  $\chi^2$ -Tests. Es handelt sich dabei um einen Hypothesentest, dessen Nullhypothese die Normalverteilungsannahme ist. Hierfür werden die Hilfsgrößen

$$\chi_{i;\text{obs}}^2 = \begin{cases} \frac{(x_i - x_{\text{ref};B})^2}{u_i^2} & \text{Teilnehmer in Gruppe „r“}, \\ 0 & \text{Teilnehmer in Gruppe „n“}. \end{cases} \quad (4.6)$$

und

$$\begin{aligned} \chi_{\text{obs}}^2 &= \sum_{i=1}^N \chi_{i;\text{obs}}^2 \\ &= \sum_{i=1}^{N_r} \frac{(x_i - x_{\text{ref};B})^2}{u_i^2} \end{aligned} \quad (4.7)$$

definiert. Der  $\chi^2$ -Test gilt als bestanden, d.h. die Nullhypothese wird nicht verworfen, wenn

$$\int_0^{\chi_{\text{obs}}^2} p_{N_r-1}(x) dx < 1 - \alpha \quad \text{mit } \alpha = 0,05. \quad (4.8)$$

Hierbei ist  $p_{N_r-1}(x)$  die Wahrscheinlichkeitsdichte der  $\chi^2$ -Verteilung mit  $N_r - 1$  Freiheitsgraden. Gewählt wird entsprechend der Normalverteilungsannahme ein Vertrauensniveau von 95 %. Wird Gleichung (4.8) nicht bestanden, so ist zunächst der Teilnehmer  $i$  mit dem größten Beitrag  $\chi_{i;\text{obs}}^2$  gemäß Gleichung (4.6) aus der Berechnung zu entfernen, d.h. dieser Teilnehmer  $i$  wird von Gruppe „r“ in Gruppe „n“ verschoben. Mit den verbleibenden  $N_r - 1$  Teilnehmern aus Gruppe „r“ werden nun gemäß den Gleichungen (4.5) und (4.3) neue Referenzdaten berechnet und der  $\chi^2$ -Test erneut durchgeführt. Diese Prozedur wird sukzessive solange durchlaufen, bis der  $\chi^2$ -Test bestanden ist.

An dieser Stelle sei angemerkt, dass für Teilnehmer die nach bestandenem  $\chi^2$ -Test in Gruppe „n“ sind, d.h. nicht in der Berechnung der Referenzdaten berücksichtigt sind, nach wie vor eine Leistungsbewertung möglich ist. Lediglich die Korrelation mit den Referenzdaten entfällt.

#### 4.1.2 Leistungsbewertung

Gemäß Abschnitt 4.1 ist klar, dass Teilnehmer aus Gruppe „r“, die in die Bestimmung der Referenzdaten einfließen, mit diesen korreliert sind. Entsprechend ist nach Abschnitt 2.2.4 diese Korrelation im  $E_{n;i}$ -

Wert zu berücksichtigen. Konkret ist in Methode B die Kovarianz aus Gleichung (2.35) gegeben durch

$$u(x_i; x_{\text{ref};B}) = \begin{cases} u_{\text{ref};B} & \text{Teilnehmer in Gruppe „r“} \\ 0 & \text{Teilnehmer in Gruppe „n“} \end{cases} \quad (4.9)$$

und ist für alle Teilnehmer in Gruppe „r“ gleich. Für den Korrelationskoeffizienten gilt nach Gleichung (2.38)

$$r_{i;\text{ref};B} = \begin{cases} \frac{u_{\text{ref};B}}{u_i} = \sqrt{w_{i;B}} & \text{Teilnehmer in Gruppe „r“} \\ 0 & \text{Teilnehmer in Gruppe „n“} \end{cases} \quad (4.10)$$

Entsprechend gilt für den  $E_{n;i}$ -Wert

$$E_{n;i} = \begin{cases} \frac{x_i - x_{\text{ref};B}}{2\sqrt{u_i^2 - u_{\text{ref};B}^2}} & \text{Teilnehmer in Gruppe „r“} \\ \frac{x_i - x_{\text{ref};B}}{2\sqrt{u_i^2 + u_{\text{ref};B}^2}} & \text{Teilnehmer in Gruppe „n“} \end{cases} \quad (4.11)$$

Für die Teilnehmer aus Gruppe „n“ ist die Leistungsbewertung im Wesentlichen ähnlich zu Methode A. Hingegen ändert sich für die Teilnehmer aus Gruppe „r“ das Vorzeichen zwischen Teilnehmer- und Referenzunsicherheit unter der Wurzel. Da  $u_i^2 - u_{\text{ref};B}^2 < u_i^2 + u_{\text{ref};B}^2$ , würde die Vernachlässigung der Korrelation für die Teilnehmer aus Gruppe „r“ zu einer Unterschätzung deren  $E_{n;i}$ -Wertes führen, d.h. ihre Leistung würde zu gut bewertet werden. Hierauf wird in Abschnitt 4.2.3 noch eingegangen. Gewissermaßen kompensiert die Berücksichtigung der Korrelation den „Vorteil“ eines Teilnehmers, den dieser dadurch hat, dass er auf die Bestimmung der Referenzdaten Einfluss nehmen kann.

## 4.2 Charakteristik der Methode

In diesem Abschnitt werden anhand ausgewählter Beispiele und Grenzfälle die Eigenschaften von Methode B dargestellt und erläutert.

### 4.2.1 Eigenschaften der Referenzunsicherheit

Zunächst wird auf die Charakteristik der Referenzunsicherheit eingegangen. Bei Betrachtung von Gleichung (4.11) fällt auf, dass der  $E_{n;i}$ -Wert für die Teilnehmer aus Gruppe „r“ nur dann definiert ist, wenn gilt

$$u_{\text{ref};B}^2 < u_i^2 \quad \forall i \in \text{Gruppe „r“}. \quad (4.12)$$

Gleichzeitig konnte jedoch in Abschnitt 2.2.4 allgemein gezeigt werden, dass der  $E_{n;i}$ -Wert immer definiert ist. Folglich handelt es sich bei Gleichung (4.12) nicht nur um eine Bedingung, sondern um einen gültigen Zusammenhang. Die nach Methode B resultierende Referenzunsicherheit ist immer kleiner als die kleinste Teilnehmerunsicherheit  $u_i$ , die in die Bestimmung der Referenzdaten einfließt. Dieser, in Gleichung (4.12) gezeigte Zusammenhang, kann auch allgemein bewiesen werden. Hierfür werden zunächst zwei Teilnehmer angenommen, aus denen die Referenzdaten ermittelt werden, wobei oBdA  $u_1 < u_2$  sei. Mit Gleichung (4.3) gilt dann

$$u_{\text{ref};B}^2 = \frac{u_1^2 u_2^2}{u_1^2 + u_2^2} < \frac{u_1^2 u_2^2}{u_2^2} = u_1^2. \quad (4.13)$$

Für  $N_r > 2$  Teilnehmer folgt der Beweis induktiv.

In Kapitel 3 wurde aus DIN EN ISO/IEC 17043:2023 zitiert, dass das Referenzlabor idealerweise so zu wählen ist, dass die resultierende Referenzunsicherheit kleiner ist die (erwarteten) Unsicherheiten des Teilnehmerkreises. Andernfalls ist die Leistungsbewertung eines Teilnehmers, dessen Unsicherheit größer ist als die Referenzunsicherheit wenig aussagekräftig. Aufgrund des Zusammenhangs aus Gleichung (4.12), kann Methode B diese Unzulänglichkeit beheben. Allerdings ist hier Vorsicht geboten, da Methode B durch den Zusammenhang in Gleichung (4.12) anfällig ist für unterschätzte Teilnehmerunsicherheiten, bzw. Ausreißer mit kleinen Unsicherheiten. Auf diese Eigenschaft wird in Abschnitt 4.2.2 eingegangen.

Eine weitere Eigenschaft der Referenzunsicherheit zeigt sich, wenn angenommen wird, dass alle  $N_r$  Teilnehmer eine Unsicherheit in der gleichen Größenordnung berichten, d.h. wenn  $u_i \simeq u_1 \quad \forall i \in [1; N_r]$ . In diesem Grenzfall gilt

$$u_{\text{ref};B} = \frac{u_1}{\sqrt{N_r}}, \quad (4.14)$$

d.h. die Referenzunsicherheit wird umso kleiner, je mehr Teilnehmerdaten in der Bestimmung der Referenzdaten berücksichtigt werden. Der in Gleichung (4.14) gezeigte Zusammenhang erinnert an den Grenzfall des arithmetischen Mittelwertes. Dieser wird im Abschnitt 4.2.2 noch einmal aufgegriffen und näher diskutiert.

Die Tatsache, dass die Referenzunsicherheit in Methode B kleiner wird, je größer die Zahl der Teilnehmer in Gruppe „r“ ist, kann erneut induktiv bewiesen werden. Als Beispiel sei an dieser Stelle der Fall  $N_r = 2$  und  $N_r = 3$  gezeigt. Für  $N_r = 2$  und oBdA  $u_1 < u_2$  gilt, wie in Gleichung (4.13) gezeigt

$$u_{\text{ref};B;N_r=2}^2 = \frac{u_1^2 u_2^2}{u_1^2 + u_2^2} < u_1^2. \quad (4.15)$$

Nun wird ein dritter Teilnehmer hinzugefügt und oBdA gilt auch hier  $u_1 < u_3$ . Für die Referenzunsicherheit gilt nun

$$u_{\text{ref};B;N_r=3}^2 = \frac{u_1^2 u_2^2 u_3^2}{u_1^2 u_2^2 + u_1^2 u_3^2 + u_2^2 u_3^2} = \frac{u_{\text{ref};B;N_r=2}^2 u_3^2}{u_{\text{ref};B;N_r=2}^2 + u_3^2} < u_{\text{ref};B;N_r=2}^2 < u_1^2. \quad (4.16)$$

Damit ist klar, dass jeder zusätzliche Teilnehmer in Gruppe „r“ die Referenzunsicherheit verkleinert, unabhängig davon, wie groß die Unsicherheit des zusätzlichen Teilnehmers ist.

Zusammengefasst kann festgehalten werden, dass die Referenzunsicherheit umso kleiner ist, je mehr Teilnehmerdaten in ihrer Bestimmung berücksichtigt sind und, je kleiner die Teilnehmerunsicherheiten sind. Die Lage der Messabweichungen und insbesondere deren Streuung bleibt in der Referenzunsicherheit unberücksichtigt.

#### 4.2.2 Eigenschaften des Referenzwertes

Im Gegensatz zur Referenzunsicherheit sind die Messabweichungen im Referenzwert berücksichtigt und die Eigenschaften des Referenzwertes ergeben sich aus dem Zusammenspiel zwischen den Teilnehmerunsicherheit und den Teilnehmerwerten.

Zu Beginn wird an dieser Stelle der Grenzfall aufgegriffen, dass alle Teilnehmer in Gruppe „r“ eine Unsicherheit in einer ähnlichen Größenordnung berichten, d.h., dass  $u_i \simeq u_1 \quad \forall i \in [1; N_r]$ . In diesem Fall gilt nach Gleichung (4.14)

$$u_{\text{ref};B} = \frac{u_1}{\sqrt{N_r}}. \quad (4.17)$$

Für die Gewichtungsfaktoren  $w_{i;B}$  folgt nach Gleichung (4.4)

$$w_{i;B} = \begin{cases} \frac{1}{N_r} & \text{Teilnehmer in Gruppe „r“,} \\ 0 & \text{Teilnehmer in Gruppe „n“,} \end{cases} \quad (4.18)$$

d.h. es wird der Gewichtungsfaktor des arithmetischen Mittelwertes, bei dem alle Messabweichungen  $x_i$  gleich gewichtet werden, reproduziert. Der Referenzwert nach Gleichung (4.5) ist in diesem Grenzfall identisch zum arithmetischen Mittelwert

$$x_{\text{ref};B} = \frac{1}{N_r} \sum_{i=1}^{N_r} x_i. \quad (4.19)$$

Der Vergleich von Gleichung (4.17) und Gleichung (4.19) zeigt, dass die Referenzunsicherheit dem Ergebnis entspricht, das nach der Fehlerfortpflanzung für den arithmetischen Mittelwert erwartet wird. Abbildung 4.1 zeigt am Beispiel von  $N_r = 2$  Teilnehmer, dass bei identischen Teilnehmerunsicherheiten der arithmetische Mittelwert resultiert, d.h. ein Wert, der exakt in der Mitte der beiden Teilnehmerwerte liegt.

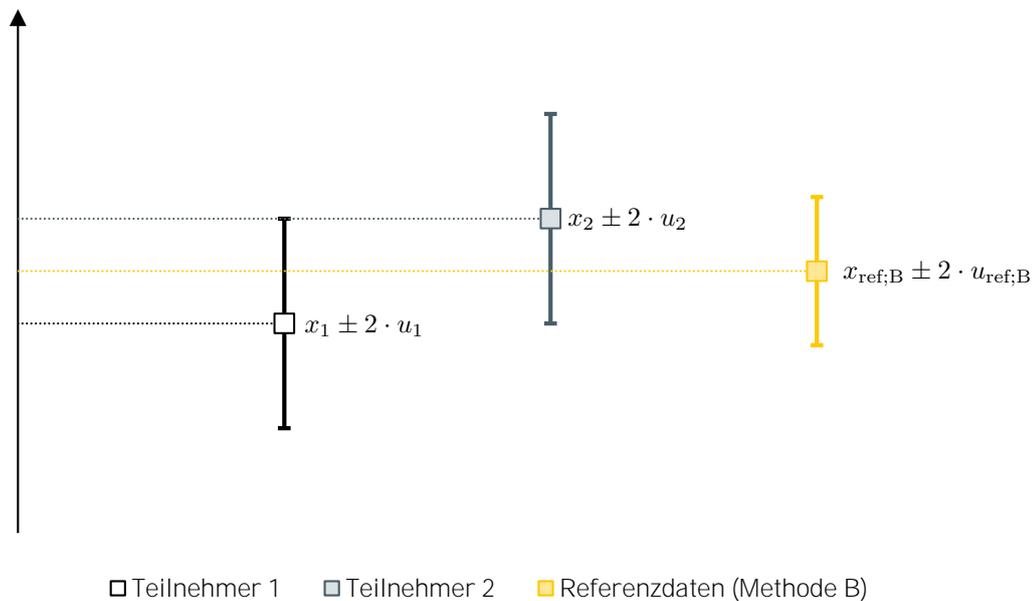
Im nächsten Schritt wird das Beispiel von  $N_r = 2$  aus Abbildung 4.1 dahingehend modifiziert, dass die Unsicherheit von Teilnehmer 1 nun deutlich kleiner sei als die Unsicherheit von Teilnehmer 2, d.h.  $u_1 \ll u_2$ . In diesem Fall kann analytisch gezeigt werden, dass nach Gleichung (4.3) für die Referenzunsicherheit gilt

$$u_{\text{ref};B}^2 = \frac{u_1^2 u_2^2}{u_1^2 + u_2^2} \simeq \frac{u_1^2 u_2^2}{u_2^2} = u_1^2, \quad (4.20)$$

d.h. die Referenzunsicherheit ist im Wesentlichen durch die Unsicherheit von Teilnehmer 1 gegeben und, wie in Gleichung (4.13) gezeigt, sogar etwas kleiner als diese. Für den Referenzwert folgt damit nach Gleichung (4.5)

$$x_{\text{ref};B} = \frac{u_{\text{ref};B}^2}{u_1^2} x_1 + \frac{u_{\text{ref};B}^2}{u_2^2} x_2 \simeq x_1 + \frac{u_1^2}{u_2^2} x_2 \simeq x_1, \quad (4.21)$$

wobei verwendet wurde, dass  $u_{\text{ref};B} \simeq u_1$  und  $u_1 \ll u_2$ , sodass  $u_1/u_2 \ll 1$ . Für den Fall, dass ein Teilnehmer eine deutlich kleinere Unsicherheit berichtet als der andere Teilnehmer, sind die Referenzdaten im



**Abbildung 4.1**

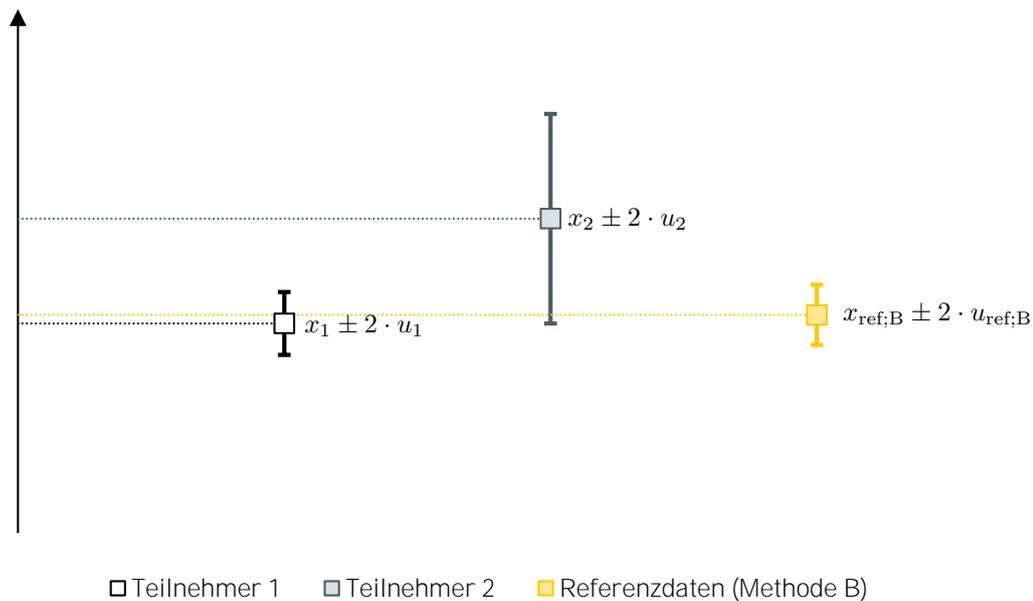
Am Beispiel von  $N_r = 2$  Teilnehmern, die in der Berechnung der Referenzdaten berücksichtigt werden, zeigt sich, dass der arithmetische Mittelwert reproduziert wird, falls beide Teilnehmer die gleiche Unsicherheit berichten. Der Referenzwert liegt exakt zwischen den beiden Teilnehmerwerten und die Referenzunsicherheit ist um den Faktor  $\sqrt{2}$  kleiner als die Teilnehmerunsicherheiten. Im gezeigten Beispiel wurden die Daten  $x_1 = 1$  und  $x_2 = 1,5$  mit  $u_1 = u_2 = 0,25$  gewählt. Für dieses Wertepaar wird der  $\chi^2$ -Test bestanden und  $x_{\text{ref};B} = 1,25$  mit  $u_{\text{ref};B} = 0,25/\sqrt{2}$ .

Wesentlichen durch die Werte des Teilnehmers mit der kleinen Unsicherheit gegeben. Der Teilnehmer mit der kleinen Unsicherheit wird durch diese stark gewichtet, wohingegen der Teilnehmer mit der größeren Unsicherheit nur eine schwache Gewichtung erhält.

$$w_{i;B} = \begin{cases} \frac{u_{\text{ref};B}^2}{u_1^2} \simeq 1 & \text{Teilnehmer } i = 1 \\ \frac{u_{\text{ref};B}^2}{u_2^2} \simeq \frac{u_1^2}{u_2^2} \ll 1 & \text{Teilnehmer } i = 2 \end{cases} \quad (4.22)$$

Dieser Grenzfall erinnert an Methode A aus Kapitel 3, da letztlich die Daten eines Teilnehmers die Referenzdaten bestimmen. Abbildung 4.2 veranschaulicht dieses Beispiel. Bildlich gesprochen zieht der Teilnehmer mit der kleinen Unsicherheit, aufgrund seines großen Gewichts, den Referenzwert zu sich heran.

Durch die Ausführungen der obigen Abschnitte, sowie das in Abbildung 4.2 gezeigte Verhalten wird klar, dass Methode B besonders anfällig ist für Teilnehmerunsicherheiten, die im Verhältnis zu den anderen Teilnehmern klein ist. Dies wird insbesondere dann kritisch, wenn der Teilnehmer mit der kleinen Unsicherheit einen Wert berichtet, der sich von allen anderen Teilnehmer deutlich unterscheidet. Im nachfolgenden wird dieser Fall eines Ausreißers mit kleiner Unsicherheit betrachtet. Dieser sei durch Teilnehmer 1 mit  $x_1 \pm u_1$  gegeben. Die restlichen Teilnehmer in Gruppe „r“ berichten der Einfachheit halber sowohl den gleichen Wert, als auch die gleiche Unsicherheit, d.h.  $x_i = x_2 \forall i \in [2; N_r]$  und  $u_i = u_2 \forall i \in [2; N_r]$ . Für die Referenzunsicherheit erhält man nach Gleichung (4.3)



**Abbildung 4.2**

Gezeigt ist das Beispiel von  $N_r = 2$  Teilnehmern, wobei die Unsicherheit von Teilnehmer 1 deutlich kleiner ist als die Unsicherheit von Teilnehmer 2. Letztere wird daher im Vergleich zu Teilnehmer 1 nur schwach gewichtet und hat kaum einen Einfluss auf die Referenzdaten. Diese sind maßgeblich durch Teilnehmer 1 bestimmt. Im gezeigten Beispiel wurden die Daten  $x_1 = 1$  und  $x_2 = 1,5$  mit  $u_1 = 0,075$  und  $u_2 = 0,25$  gewählt. Für dieses Wertepaar wird der  $\chi^2$ -Test bestanden und  $x_{\text{ref;B}} = 1,04$  mit  $u_{\text{ref;B}} = 0,072$ .

$$u_{\text{ref;B}}^2 = \frac{u_1^2 u_2^2}{u_2^2 + (N_r - 1) u_1^2} \quad (4.23)$$

und für den Referenzwert gemäß Gleichung (4.5)

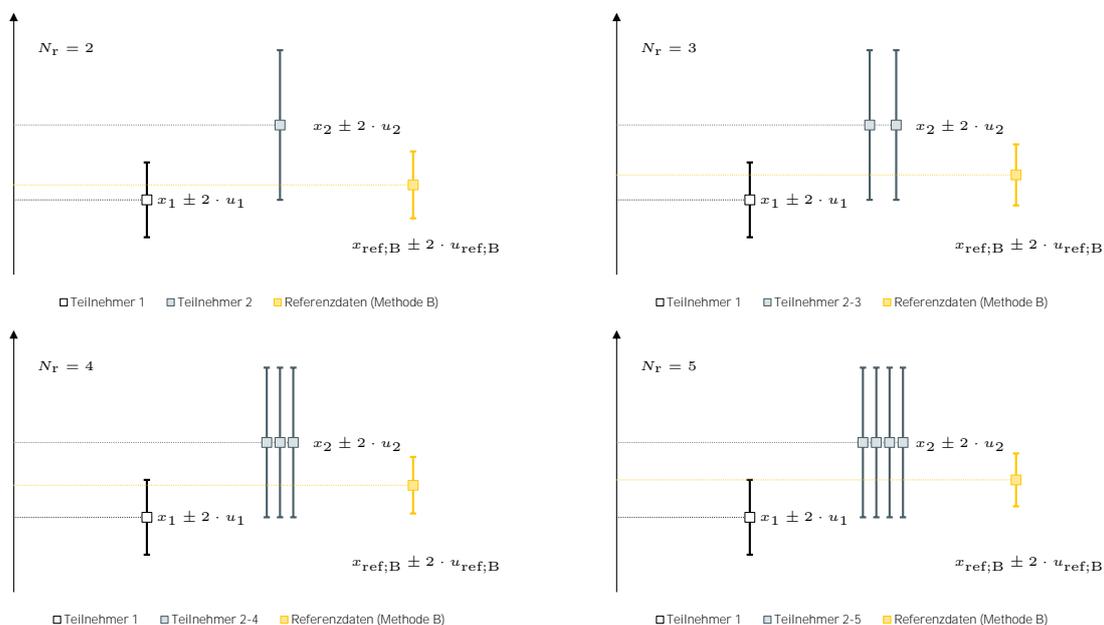
$$x_{\text{ref;B}} = \frac{1}{u_2^2 + (N_r - 1) u_1^2} (u_2^2 x_1 + (N_r - 1) u_1^2 x_2). \quad (4.24)$$

Die  $N_r - 1$  Teilnehmer stellen zusammen ein Gegengewicht zu der kleinen Unsicherheit von Teilnehmer 1 dar. Dieses Gegengewicht ist umso stärker, je größer  $N_r$  ist. Zudem hängt es vom konkreten Verhältnis der Unsicherheiten  $u_1$  und  $u_2$  ab. Wie Abbildung 4.3 zeigt, ist für  $N_r = 2$  der Referenzwert noch maßgeblich von Teilnehmer 1 dominiert. Je mehr Teilnehmer jedoch hinzugefügt werden, welche die gleichen Werte aufweisen wie Teilnehmer 2, umso mehr wird der Referenzwert von dieser Gruppe angezogen. Ein Gleichgewicht der beiden Gewichte entsteht dann, wenn

$$N_r = \frac{u_2^2}{u_1^2} + 1. \quad (4.25)$$

In diesem Fall ist der Referenzwert durch einen Wert in der Mitte zwischen Teilnehmer 1 und den restlichen Teilnehmern gegeben. Weitere Teilnehmer mit den Werten  $x_2 \pm u_2$  führen nun dazu, dass der Referenzwert immer weniger durch den Wert von Teilnehmer 1 beeinflusst ist. Das bedeutet, dass das Gewicht einer kleinen Teilnehmerunsicherheit durch hinreichend viele Teilnehmer mit ähnlichen Werten kompensiert werden kann. Allerdings ist auch klar, dass Methode B zunächst anfällig ist für

Teilnehmer mit kleinen Unsicherheiten, insbesondere bei geringer Datenlage. Kritisch wird dies, wenn ein Teilnehmer die berichtete Unsicherheit unterschätzt. Ein Hinweis dafür kann sein, dass es eine Diskrepanz zwischen der Messwertstreuung aller Teilnehmer und der berichteten Teilnehmerunsicherheit gibt. Je nach Datenlage ist Methode B somit anfällig und nicht robust gegen Ausreißer mit kleinen Unsicherheiten. Der  $\chi^2$ -Test bietet die Möglichkeit solche Ausreißer zu erkennen und von Gruppe „r“ in Gruppe „n“ zu verschieben. Allerdings fängt dieser Mechanismus je nach Datenlage nicht zwangsläufig jeden Ausreißer mit kleiner Unsicherheit ab. In den gezeigten Beispielen von Abbildung 4.3 ist der  $\chi^2$ -Test immer bestanden.



**Abbildung 4.3**

Gezeigt ist das Datenbeispiel  $x_1 = 1$  mit  $u_1 = 0,25$  und  $x_2 = 2$  mit  $u_2 = 0,5$ . Die Teilnehmerzahl  $N_r$ , die in der Bestimmung der Referenzdaten berücksichtigt ist, wird sukzessive von links nach rechts und von oben nach unten erhöht. Der  $\chi^2$ -Test ist in allen vier Fällen bestanden. Zu Beginn ist der Einfluss des Teilnehmers 1, d.h. des Teilnehmer mit der kleinen Unsicherheit noch signifikant. Dieser wird durch das Gegengewicht der restlichen Teilnehmer und das sukzessive Hinzufügen weiterer Teilnehmer mit den Werten  $x_2 \pm u_2$  reduziert. Im gezeigten Datenbeispiel ist das Gleichgewicht nach Gleichung (4.25) für  $N_r = 5$  Teilnehmer erreicht und ein Referenzwert exakt zwischen  $x_1$  und  $x_2$  resultiert.

### 4.2.3 Eigenschaften der Leistungsbewertung

In diesem Abschnitt wird auf das Verhalten des  $E_{n;i}$ -Wertes eingegangen werden, für Teilnehmer  $i$ , die in die Bestimmung der Referenzdaten einfließen, d.h. in Gruppe „r“ sind. Einige Merkmale wurden bereits im obigen Abschnitt angesprochen (vgl. Abbildung 4.5) und sollen hier noch einmal näher betrachtet werden. Wie in Gleichung (4.11) gezeigt, führt die Korrelation zwischen Teilnehmer- und Referenzwert dazu, dass der  $E_{n;i}$ -Wert sich ändert zu

$$E_{n;i} = \frac{x_i - x_{\text{ref};B}}{2\sqrt{u_i^2 - u_{\text{ref};B}^2}} \quad \text{für } i \text{ in Gruppe „r“}. \quad (4.26)$$

Wie schon in Abschnitt 4.1.2 geschrieben, wird durch die Berücksichtigung der Korrelation der Vorteil zum Teil kompensiert, den ein Teilnehmer hat, wenn dieser auf die Referenzdaten Einfluss nehmen kann. Im Fall von Methode B zeigt sich dies konkret an der Subtraktion der Varianzen im Nenner.

Von Interesse ist insbesondere der Fall, dass ein Teilnehmer (oBdA  $i = 1$ ) die Referenzdaten maßgeblich definiert. In diesem Fall ist  $x_{\text{ref};B} \simeq x_1$ . Gleichzeitig ist aber auch  $u_{\text{ref};B} \simeq u_1$ , sodass nicht nur der Zähler von Gleichung 4.26 sehr klein wird sondern auch der Nenner. An dieser Stelle sei nochmal angemerkt, dass der Nenner stets definiert ist, da  $u_{\text{ref};B} < u_i \forall i \in [1; N_r]$ . Betrachtet wird nun Gleichung (4.26) für Teilnehmer  $i = 1$ , wobei der Referenzwert umgeschrieben wird in

$$x_{\text{ref};B} = u_{\text{ref};B}^2 \sum_{i=1}^{N_r} \frac{x_i}{u_i^2} = \frac{u_{\text{ref};B}^2}{u_1^2} x_1 + u_{\text{ref};B}^2 \sum_{i=2}^{N_r} \frac{x_i}{u_i^2}, \quad (4.27)$$

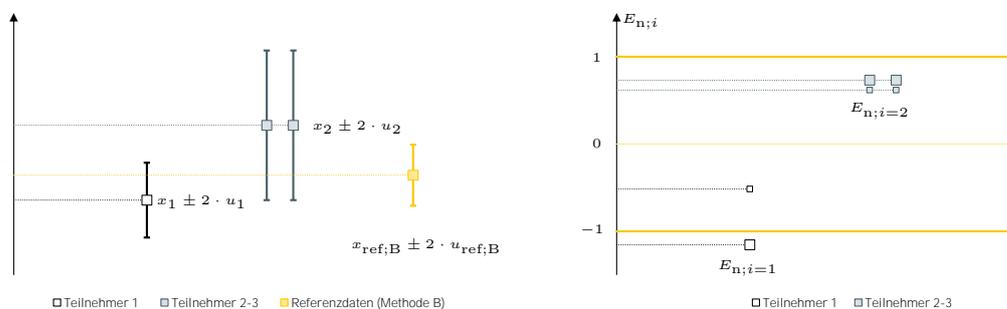
sodass

$$\begin{aligned} E_{n;i=1} &= \frac{x_1 \left(1 - \frac{u_{\text{ref};B}^2}{u_1^2}\right) - u_{\text{ref};B}^2 \sum_{i=2}^{N_r} \frac{x_i}{u_i^2}}{2u_1^2 \sqrt{1 - \frac{u_{\text{ref};B}^2}{u_1^2}}} \\ &= \frac{x_1}{2u_1^2} \sqrt{1 - \frac{u_{\text{ref};B}^2}{u_1^2}} - \frac{u_{\text{ref};B}^2 \sum_{i=2}^{N_r} \frac{x_i}{u_i^2}}{2u_1^2 \sqrt{1 - \frac{u_{\text{ref};B}^2}{u_1^2}}} \\ &\stackrel{u_{\text{ref};B} \simeq u_1}{\simeq} - \frac{u_{\text{ref};B}^2 \sum_{i=2}^{N_r} \frac{x_i}{u_i^2}}{2u_1^2 \sqrt{1 - \frac{u_{\text{ref};B}^2}{u_1^2}}}. \end{aligned} \quad (4.28)$$

Mit dieser Abschätzung zeigt sich, dass die Konvergenz zu Null im Nenner dominiert, sodass ein  $E_{n;i=1}$ -Wert nahe Null nicht zu erwarten ist, trotz bzw. besser gerade wegen des starken Einflusses von Teilnehmer 1 auf den Referenzwert. Hier zeigt sich insbesondere, dass das Vernachlässigen der Korrelation zu einer signifikanten Fehlbewertung führen kann. Denn ohne die Korrelation würden die beiden Varianzen im Nenner von Gleichung (4.26) addiert. Für  $x_{\text{ref};B} \simeq x_1$  würde dann nur der Zähler gegen Null konvergieren, was zu einer zu positiven Leistungsbewertung führen würde. Durch die Berücksichtigung der Korrelation hingegen besteht die Möglichkeit, dass ein  $E_{n;i=1}$ -Wert betragsmäßig nahe oder sogar größer als 1 resultiert, obwohl  $x_{\text{ref};B} \simeq x_1$ .

Um das gezeigte analytische Verhalten zu veranschaulichen, wird an dieser Stelle das Beispiel aus Abbildung 4.3 für  $N_r = 3$  Teilnehmer aufgegriffen. Teilnehmer 1 beeinflusst den Referenzwert maßgeblich. Dennoch resultiert ein  $E_{n;i=1}$ -Wert, der betragsmäßig größer ist als eins, wie in Abbildung 4.4 gezeigt. Obwohl der Referenzwert innerhalb des 95 %-igen Vertrauensbereichs von Teilnehmer 1 liegt, bzw. der Wert von Teilnehmer 1 innerhalb des 95 %-igen Vertrauensbereichs des Referenzwertes liegt, resultiert ein  $E_{n;i=1}$ -Wert, der betragsmäßig größer ist als eins. Hier wird erneut deutlich, dass die graphische Überlappung der Unsicherheitsbalken nicht mit der positiven Leistungsbewertung durch

den  $E_{n;i}$ -Wert gleichzusetzen ist. Insbesondere, da in einer graphischen Überlappung keine Korrelationen berücksichtigt werden. In Abbildung 4.4 ist zudem ersichtlich, dass die Berücksichtigung der Korrelation für Teilnehmer 1 zu einer signifikanten Änderung des  $E_{n;i}$ -Wertes führt, wohingegen sich der  $E_{n;i}$ -Wert für Teilnehmer 2 nur minimal verschlechtert. Der Grund dafür ist, dass Teilnehmer 1 mit dem Referenzwert stärker korreliert ist, was anhand des Korrelationskoeffizienten aus Gleichung (4.10) ersichtlich ist. Damit wird noch einmal deutlich, dass die Berücksichtigung der Korrelation den Vorteil eines Teilnehmers kompensiert, den dieser zunächst hat, wenn er aufgrund einer kleinen Unsicherheit einen signifikanten Einfluss auf den Referenzwert hat. Die gezeigten Ausführungen machen deutlich, dass die Berücksichtigung der Korrelation keine Option ist, sondern zwingend erforderlich ist für eine korrekte Leistungsbewertung. Da im vorliegenden Fall die Korrelation zwischen Teilnehmer- und Referenzwert analytisch bekannt ist, ist die Berücksichtigung jederzeit möglich.

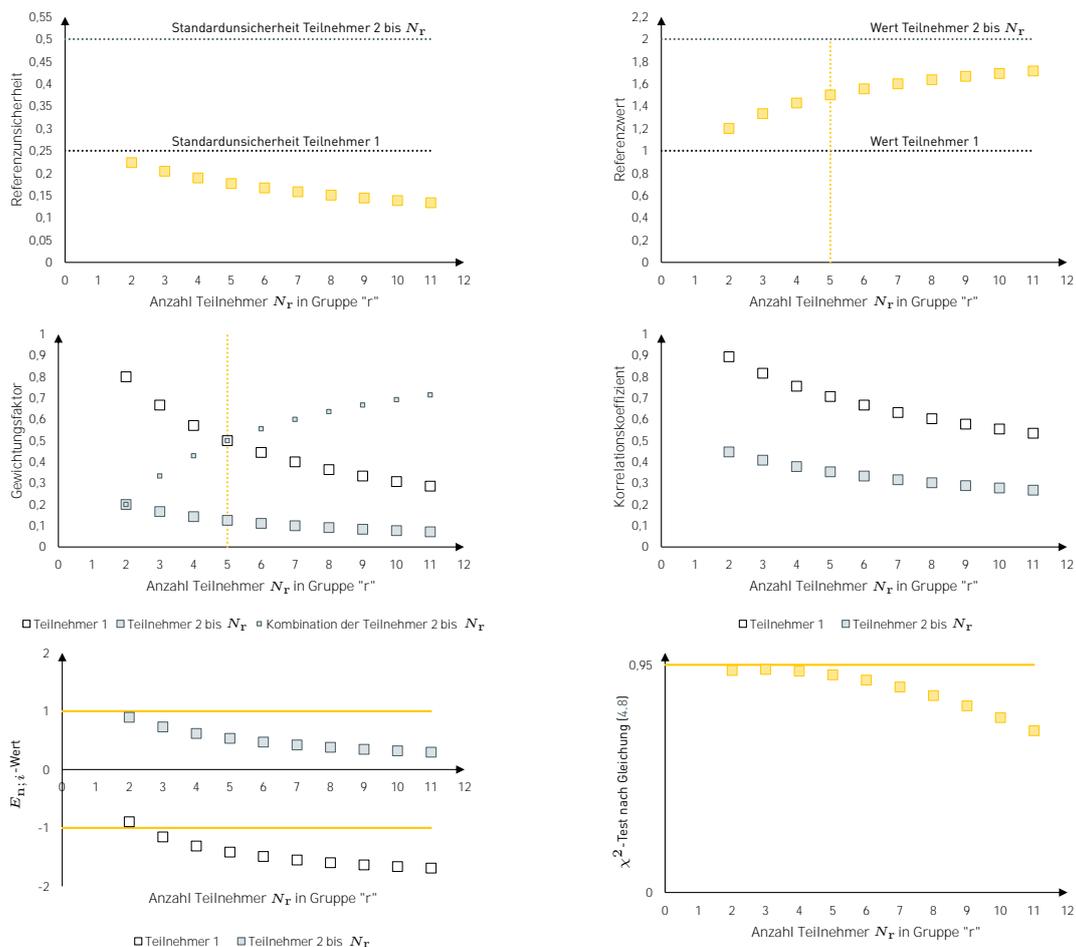


**Abbildung 4.4**

Am Beispiel aus Abbildung 4.3 für  $N_r = 3$  wird deutlich, dass durch die Berücksichtigung der Korrelation im  $E_{n;i}$ -Wert der Vorteil, den ein Teilnehmer hat, wenn er auf die Bestimmung des Referenzwertes Einfluss hat, kompensiert wird. Die großen Quadrate zeigen die richtige Berechnung des  $E_{n;i}$ -Wertes, d.h. mit Berücksichtigung der Korrelation. Die kleinen Quadrate zeigen die  $E_{n;i}$ -Werte bei vernachlässigter Korrelation. Im vorliegenden Datenbeispiel ist  $E_{n;i=1} \simeq -1,15$  und  $E_{n;i=2} \simeq 0,73$ , jeweils mit Korrelation. Das heißt, obwohl Teilnehmer 1 einen signifikanten Einfluss auf den Referenzwert hat, ist sein  $E_{n;i}$ -Wert betragsmäßig größer als eins. Die Berücksichtigung der Korrelation führt für diesen Teilnehmer zu einer signifikanten Änderung des  $E_{n;i}$ -Wertes (vgl. ohne Korrelation:  $E_{n;i=1} \simeq -0,52$  und  $E_{n;i=2} \simeq 0,62$ ). Da Teilnehmer 2 und 3 mit dem Referenzwert weniger stark korreliert sind, führt die Berücksichtigung der Korrelation hier zu einer kleineren Änderung.

Abschließend ist in Abbildung 4.5 die Abhängigkeit verschiedener Parameter von der Teilnehmerzahl  $N_r$  gezeigt, wobei das Wertebeispiel aus Abbildung 4.3 aufgegriffen wird. Teilnehmer 1 berichtet als einziger die Werte  $x_1 = 1$  und  $u_1 = 0,25$ . Alle restlichen Teilnehmer  $i \in [2; N_r]$  berichten die gleichen Werte  $x_2 = 2$  und  $u_2 = 0,5$ . Die Gesamtzahl  $N_r$  wird sukzessive erhöht, sodass immer mehr Teilnehmer vorliegen, welche die Werte  $x_2$  und  $u_2$  berichten. Die in Abschnitt 4.2.1 diskutierte Eigenschaft der Referenzunsicherheit, dass diese mit steigender Teilnehmerzahl immer kleiner wird, ist in Abbildung 4.5 zu erkennen. Außerdem wird noch einmal deutlich, dass die Referenzunsicherheit in jedem Fall kleiner ist als die kleinste Teilnehmerunsicherheit, die in die Berechnung einfließt. Hinsichtlich des Referenzwertes zeigt sich zunächst das in Abschnitt 2.2.4 allgemein beschriebene Verhalten, dass der Referenzwert immer zwischen dem minimalen und dem maximalen Teilnehmerwert liegt. Zudem zeigt sich an diesem Beispiel, dass für kleine Teilnehmerzahlen  $N_r$  der Referenzwert maßgeblich durch Teilnehmer 1 dominiert ist, da dieser aufgrund seiner verhältnismäßig kleinen Unsicherheit eine starke Gewichtung erhält. Dieses Gewicht der kleinen Unsicherheit wird reduziert, durch die Anzahl  $N_r - 1$  der

Teilnehmer, die den Wert von Teilnehmer 2 berichten. Damit wird noch einmal unterstrichen, dass das Gegengewicht zu einer kleinen Unsicherheit durch die Gesamtzahl der restlichen Teilnehmer gegeben ist. Im gewählten Beispiel ist ein Gleichgewicht bei  $N_r = 5$  erreicht und der Referenzwert befindet sich exakt zwischen  $x_1$  und  $x_2$ . Das graphische Ergebnis stimmt mit dem Wert gemäß Gleichung (4.25) überein. Das Gleichgewicht ist auch in den Gewichtungsfaktoren ersichtlich, allerdings nur, wenn für die Teilnehmer  $i \in [2; N_r]$  das kombinierte Gewicht betrachtet wird. Dieses steigt mit zunehmender Teilnehmerzahl  $N_r$  an. Das Gewicht eines einzelnen Teilnehmer sinkt jedoch mit zunehmender Teilnehmerzahl. Dieses Verhalten ist für alle Teilnehmer  $i \in [1; N_r]$  zu beobachten und zu erwarten.



**Abbildung 4.5**

Für das Datenbeispiel aus Abbildung 4.3 ist die Abhängigkeit von Referenzunsicherheit (oben links), Referenzwert (oben rechts), Gewichtungsfaktoren (mitte links), Kovarianz (mitte rechts),  $E_{n,i}$ -Wert (unten links) und  $\chi^2$ -Test (unten rechts) gezeigt.

Wie in Gleichung (4.10) ersichtlich, ist der Korrelationskoeffizient eines Teilnehmers im Wesentlichen durch die Wurzel des Gewichtungsfaktors gegeben, sodass auch hier ersichtlich wird, dass die Korrelation eines einzelnen Teilnehmers mit dem Referenzwert bei steigender Teilnehmerzahl wie erwartet abnimmt. In Abbildung 4.5 wird zudem deutlich, dass die Korrelation zwischen Teilnehmer 1 und

Referenzwert zwar mit steigender Teilnehmerzahl abnimmt, dennoch im betrachteten Bereich noch relativ hoch ist. Diese Tatsache spiegelt sich auch im  $E_{n;i=1}$ -Wert wieder, der mit steigender Teilnehmerzahl betragsmäßig immer größer wird. Hierfür sind zwei Gründe aufzuführen. Zum einen weist Teilnehmer 1 nach wie vor eine Korrelation von über 50 % mit dem Referenzwert auf. Zwar taucht der Korrelationskoeffizient im  $E_{n;i}$ -Wert von Gleichung (4.11) nicht explizit auf, allerdings ist auch klar, dass Teilnehmer 1 im Vergleich zu den restlichen Teilnehmern immer den kleinsten Nenner haben wird, da  $u_1 - u_{\text{ref};B} < u_2 - u_{\text{ref};B} \quad \forall i \in [2; N_r]$ . Zum anderen ist der Referenzwert mit steigender Teilnehmerzahl immer stärker durch die Werte der Teilnehmergruppe „n“ bis  $N_r$  dominiert, sodass für  $N_r > 5$  auch der Zähler im  $E_{n;i}$ -Wert betragsmäßig größer ist als der Zähler der anderen Teilnehmer, d.h.  $|x_1 - x_{\text{ref};B}| < |x_2 - x_{\text{ref};B}| \quad \forall i \in [2; N_r > 5]$ . Die Teilnehmer 2 bis  $N_r$  hingegen profitieren, da sich sowohl der Referenzwert zu ihren Gunsten verändert und gleichzeitig auch die Korrelation eines einzelnen Teilnehmers mit dem Referenzwert abnimmt, je mehr Teilnehmer hinzugefügt werden.

Den  $E_{n;i}$ -Wert betreffend fällt zudem auf, dass für  $N_r = 2$ , Teilnehmer 1 und Teilnehmer 2 den betragsmäßig gleichen Wert erzielen. Diese Beobachtung ist für Methode B allgemeingültig und nicht von den gewählten Werten abhängig. Analytisch lässt sich diese Eigenschaft leicht zeigen, denn es gilt für  $N_r = 2$

$$u_{\text{ref};B} = \frac{u_1 u_2}{\sqrt{u_1^2 + u_2^2}} \quad (4.29)$$

und

$$x_{\text{ref};B} = \frac{1}{u_1^2 + u_2^2} (u_2^2 x_1 + u_1^2 x_2). \quad (4.30)$$

Damit folgt für den  $E_{n;i}$  Wert für Teilnehmer 1 und 2

$$\begin{aligned} E_{n;i=1} &= \frac{x_1 - x_2}{2\sqrt{u_1^2 + u_2^2}} \\ E_{n;i=2} &= \frac{x_2 - x_1}{2\sqrt{u_1^2 + u_2^2}} = -E_{n;i=1}. \end{aligned} \quad (4.31)$$

Unabhängig von den berichteten Werten der beiden Teilnehmer werden sie jeweils die gleiche Leistungsbewertung erzielen, d.h. beide Teilnehmer bestehen die Vergleichsmessung gemeinsam oder eben nicht. Dieses Ergebnis wird erneut nur dann richtig reproduziert, wenn die Korrelation berücksichtigt wird. Auf die Handhabung von Methode B bei bilateralen Eignungsprüfungen wird in Kapitel 8 noch ausführlich eingegangen.

Der Vollständigkeit halber sei hier noch angemerkt, dass in Abbildung 4.5 auch der  $\chi^2$ -Test nach Gleichung 4.8 gezeigt ist. Dieser ist in allen Fällen bestanden. Für kleine Teilnehmerzahlen wird der  $\chi^2$ -Test nur knapp bestanden. Hier zeigt sich, dass ein stärkerer Ausreißer, d.h. wenn Teilnehmer 1 entweder einen noch kleineren Wert oder eine noch kleinere Unsicherheit berichtet, zu einem Fehlschlagen des  $\chi^2$ -Tests führen würde. Insbesondere bei größerer Teilnehmerzahl zeigt sich jedoch, dass der  $\chi^2$ -Test nicht zwangsläufig fehlschlagen muss. Abschließend sei außerdem angemerkt, dass die Teilnehmer 2 bis  $N_r$  letztlich von Teilnehmer 1 profitieren. Ein Ausreißer mit kleiner Unsicherheit kann daher zu

einer Gesamtbeeinflussung der Auswertung führen, falls er durch den  $\chi^2$ -Test nicht eliminiert wird. Methode B ist daher nur begrenzt robust. Insbesondere ist hinsichtlich „moderaten“ Ausreißern mit kleiner Unsicherheit Vorsicht geboten.

Anmerkung: Der Anschaulichkeit halber wurde in diesem Kapitel u.a. das Beispiel von zwei Teilnehmern betrachtet, um die Eigenschaften der Methode B aufzuzeigen. Wie in Kapitel 8 gezeigt ist eine statistische Auswertung in diesem Fall jedoch nicht empfohlen und mit Vorsicht zu betrachten.

## 5 Methode C — Robuste Abwandlung des arithmetischen Mittelwertes (ISO 13528:2022)

In diesem Kapitel wird zunächst der arithmetische Mittelwert betrachtet. Darauf aufbauend wird der robuste Algorithmus aus ISO 13528:2022 Anhang C.3 [3] beschrieben und analysiert. Dieser Algorithmus ist nachfolgend als Methode C bezeichnet. Wie bereits in Methode B, fließen auch in Methode C, bzw. im arithmetischen Mittelwert allgemein, Teilnehmerdaten in die Bestimmung der Referenzdaten ein. Insbesondere ist der arithmetische Mittelwert durch das lineare Modell aus Abschnitt 2.2.4 beschreibbar. Für Methode C, d.h. für den robusten Algorithmus aus ISO 13528:2022, ist das lineare Modell aufgrund der einzelnen Iterationsschritte des Algorithmus nur in abgewandelter Form anwendbar. Hierauf wird in Abschnitt 5.2.2 näher eingegangen.

Bei Methode C handelt es sich um eine streuungsbasierte Auswertungsmethode mit einer vereinheitlichten Gewichtung, die sich somit grundlegend von Methode B unterscheidet. Die Charakteristik von Methode C wird in den nachfolgenden Abschnitten diskutiert. Zu Beginn wird hierfür zunächst der arithmetische Mittelwert betrachtet, der als Grundlage von Methode C verstanden werden kann. Aus diesem Grund wird der arithmetische Mittelwert nachfolgend als Methode  $C_0$  bezeichnet.

Betrachtet wird in diesem Kapitel der in 2.1 definierte Datensatz, d.h. es liegen  $N$  Teilnehmer vor, wobei die berichtete Messabweichung durch  $x_i$  und die dazugehörige erweiterte Messunsicherheit durch  $U_i$  gegeben sind. In Kapitel 2 wurde bereits die für dieses Dokument gültige Annahme getroffen, dass  $U_i = 2 \cdot u_i \forall i \in [1; N]$ , d.h., dass die einzelnen Teilnehmerergebnisse durch eine Normalverteilung beschrieben werden. Zudem wird in diesem Kapitel die Einteilung der  $N$  Teilnehmer in zwei Gruppen aus Abschnitt 2.2.4 verwendet, d.h. lediglich die  $N_r$  Teilnehmer aus Gruppe „r“ haben einen Einfluss auf den Referenzwert. Eine Vorauswahl kann, wie in Kapitel 3 dargelegt, anhand qualitativer Kriterien getroffen werden, sodass beispielsweise nur nach DIN EN ISO/IEC 17025 akkreditierte Labore oder aus einem anderen Grund ausgezeichnete Labore in die Bestimmung der Referenzdaten einfließen. Alle Teilnehmer, die Gruppe „r“ zugeordnet werden, haben in Methode  $C_0$  und C letztlich einen Einfluss auf die Referenzdaten, obgleich dieser Einfluss sehr unterschiedlich ausfallen kann, wie in den nachfolgenden Abschnitten gezeigt wird. Ein Mechanismus zur Verschiebung von Teilnehmern aus Gruppe „r“ in Gruppe „n“, d.h. in die Gruppe, deren Teilnehmer nicht in der Bestimmung der Referenzdaten berücksichtigt werden, existiert in den nachfolgend dargestellten Methoden nicht. Vielmehr setzt Methode C auf die Modifizierung, statt die Eliminierung von Ausreißern, wie Abschnitt 5.2.1 zeigt.

### 5.1 Arithmetischer Mittelwert (Methode $C_0$ )

Der arithmetische Mittelwert für Methode  $C_0$  ist wie in [17] definiert durch

$$x_{\text{ref};C_0} = \frac{1}{N_r} \sum_{i=1}^{N_r} x_i. \quad (5.1)$$

Im Vergleich zu Gleichung (2.29) zeigt sich, dass hier ein lineares Modell vorliegt, wobei die Gewichtungsfaktoren  $w_i$  definiert sind durch

$$w_{i;C_0} = \begin{cases} \frac{1}{N_r} & \text{Teilnehmer in Gruppe „r“} \\ 0 & \text{Teilnehmer in Gruppe „n“} \end{cases} \quad (5.2)$$

Der Gewichtungsfaktor ist somit für alle Teilnehmer aus Gruppe „r“ gleich und alle Werte werden einheitlich gewichtet. Die Unsicherheiten der Teilnehmer spielen bei der Bestimmung des arithmetischen Mittelwertes keine Rolle. Diese Art der Gewichtung steht im Gegensatz zu Methode B, bei der die Teilnehmer über ihre individuelle Standardunsicherheit gewichtet werden. Aus Gleichung (5.2) kann die Kovarianz  $u(x_i; x_{\text{ref};C_0})$  gemäß Gleichung (2.35) bestimmt werden. Das Ergebnis lautet

$$u(x_i; x_{\text{ref};C_0}) = \begin{cases} \frac{u_i^2}{N_r} & \text{Teilnehmer in Gruppe „r“} \\ 0 & \text{Teilnehmer in Gruppe „n“} \end{cases} \quad (5.3)$$

Der arithmetische Mittelwert an sich enthält keinen Mechanismus um Ausreißer von Gruppe „r“ in Gruppe „n“ zu verschieben, d.h. aus der Bestimmung der Referenzdaten heraus zu nehmen. Geeignete Tests auf inkonsistente Daten sind an dieser Stelle zusätzlich durchzuführen. Als Beispiel für einen statistischen Ausreißertest sei an dieser Stelle auf den Grubbs-Test [18, 19] verwiesen. Der robuste Algorithmus aus Methode C behebt dieses Defizit von Methode C<sub>0</sub>. Abschnitt 5.2.1 vorwegnehmend sei angemerkt, dass statistische Tests Ausreißer typischerweise identifizieren und dann aus den Berechnungsdaten entfernen. Der robuste Algorithmus von Methode C hingegen identifiziert Ausreißer und modifiziert deren Wert in geeigneter Weise. Ergänzend können auch qualitative Kriterien angewandt werden, sodass bspw. nur Nationalinstitute oder nach DIN EN ISO/IEC 17025:2018 akkreditierte Labore in der Bestimmung der Referenzdaten berücksichtigt werden.

Für die zum arithmetischen Mittelwert gehörende Referenzunsicherheit  $u_{\text{ref};C_0}$  gibt es zwei Möglichkeiten. Die erste ist durch die empirische Standardabweichung des arithmetischen Mittelwertes gegeben, d.h.

$$u_{\text{ref};C_0;1}^2 = \frac{1}{N_r(N_r - 1)} \sum_{i=1}^{N_r} (x_i - x_{\text{ref};C_0})^2. \quad (5.4)$$

Insbesondere bleiben die Teilnehmerunsicherheiten  $u_i$  an dieser Stelle weiter unberücksichtigt. Stattdessen wird auf die Streuung der Teilnehmerdaten zurück gegriffen. Da dieser Ansatz letztlich auch in Methode C in abgewandelter Form verwendet wird, zeigt sich hier bereits die eingangs genannte Eigenschaft, dass es sich im Gegensatz zu Methode B um eine streuungsbasierte Auswertung handelt. Diese Eigenschaft steckt letztlich bereits im arithmetischen Mittelwert aus Gleichung (5.1). An dieser Stelle sei auf Anmerkung 2 im Unterkapitel 2.2.4 verwiesen, da über Gleichung (5.4) die Rückführung der Referenzdaten nicht zwangsläufig sicher gestellt ist.

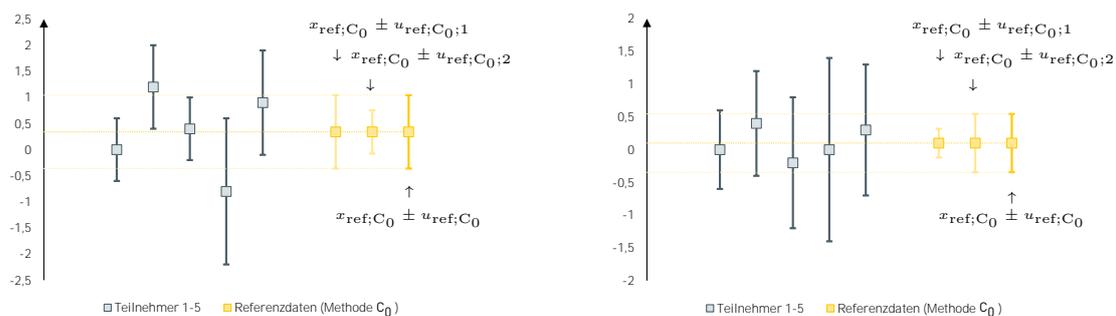
Die zweite Möglichkeit für die Referenzunsicherheit ist daher durch die Berücksichtigung der Teilnehmerunsicherheiten  $u_i$  entsprechend der Fehlerfortpflanzung nach Gleichung (2.30) aus Abschnitt 2.2.4 gegeben, d.h.

$$u_{\text{ref};C_0;2}^2 = \frac{1}{N_r^2} \sum_{i=1}^{N_r} u_i^2. \quad (5.5)$$

In diesem Fall würden die Teilnehmerunsicherheiten berücksichtigt werden, allerdings bliebe die Streuung der Teilnehmerwerte in der Referenzunsicherheit unberücksichtigt. Eine Möglichkeit die beiden Gleichungen (5.4) und (5.5) zu verknüpfen ist nach [7] gegeben durch

$$u_{\text{ref};C_0} = \max(u_{\text{ref};C_0;1}; u_{\text{ref};C_0;2}). \quad (5.6)$$

Die Anwendung dieser Gleichung ist in Abbildung 5.1 gezeigt. Dort zeigt sich, dass durch die Ergänzung von  $u_{\text{ref};C_0;2}$  verhindert wird, dass die Referenzunsicherheit unterschätzt wird, falls die Streuung der Teilnehmerwert deutlich kleiner ist als die Teilnehmerunsicherheiten. Eine Berücksichtigung der Konsistenz der Daten ist damit jedoch nicht gegeben, d.h. Methode  $C_0$  (und auch Methode C) berücksichtigen nicht, ob die Streuung im Rahmen der berichteten Unsicherheiten erklärbar ist. Diese Eigenschaft wird in Methode D (Kapitel 6) ergänzt.



**Abbildung 5.1**

Das Beispiel links zeigt den Fall, dass die Streuung der Teilnehmerwerte  $x_i$  größer ist als die Fehlerfortpflanzung der Teilnehmerunsicherheiten. Im rechten Beispiel ist die Streuung deutlich kleiner als die einzelnen Teilnehmerunsicherheiten und deren quadratische Addition. Die empirische Standardunsicherheit würde in diesem Fall die Referenzunsicherheit unterschätzen. In beiden Fällen zeigt sich, dass der Referenzwert unabhängig ist von den Teilnehmerunsicherheiten.

In Kapitel 2 wurde die Annahme getroffen, dass die Daten der einzelnen Teilnehmer durch eine Normalverteilung beschrieben werden können. Auch für die Referenzdaten aus Methode B, C und D wird folglich eine Normalverteilung angenommen. In Gleichung (2.7) wird daher der Erweiterungsfaktor  $k = 2$  verwendet. Diese Annahme stößt bei Methode  $C_0$  jedoch an ihre Grenzen, falls  $u_{\text{ref};C_0} = u_{\text{ref};C_0;1}$ . In diesem Fall wird der Referenzunsicherheit ein Freiheitsgrad von  $N_r - 1$  zugeordnet. Folglich ist der effektive Freiheitsgrad der Unsicherheit im Nenner des  $E_{n;i}$ -Wertes aus Gleichung (2.7) endlich, sodass der Erweiterungsfaktor  $k$  entsprechend der Student-t Verteilung auch größer als zwei sein kann. Die Verwendung von  $k = 2$  führt in diesem Fall zu einem größeren  $E_{n;i}$ -Wert, sodass formal die Möglichkeit besteht, dass Leistungen von Teilnehmern zu schlecht bewertet werden. Entscheidend für den effektiven Freiheitsgrad ist die Unsicherheit  $u_i$  des einzelnen Teilnehmers. Je kleiner diese Unsicherheit ist, desto kleiner ist auch der effektive Freiheitsgrad, wodurch wiederum der Erweiterungsfaktor der Student-t Verteilung größer wird. Das bedeutet, dass Teilnehmer mit einer verhältnismäßig kleinen Unsicherheit von einer Anpassung des Erweiterungsfaktors entsprechend der Student-t Verteilung eher profitieren

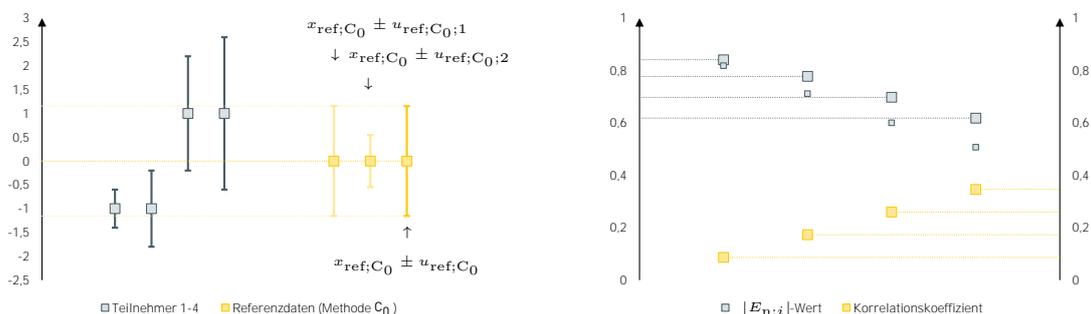
als Teilnehmer mit einer größeren Unsicherheit. Allgemein kann gezeigt werden, dass der effektive Freiheitsgrade im Intervall  $[N_r - 1; (N_r - 1)(1 + ((N_r - 2)N_r + 1)^2)]$  liegt. Der effektive Freiheitsgrad ist insbesondere dann klein, wenn  $N_r$  klein ist und/oder wenn die Streuung der Teilnehmerwerte größer ist als die Unsicherheiten der Teilnehmer, d.h. wenn eine Diskrepanz in den Daten vorliegt (s. dazu auch Methode D, Kapitel 6). Die genannte Eigenschaft, dass der Erweiterungsfaktor letztlich von der Teilnehmerunsicherheit abhängt und insbesondere Teilnehmer mit einer kleinen Unsicherheit in der Leistungsbewertung davon profitieren, führt dazu, dass die Anpassung des Erweiterungsfaktors in diesem Dokument nicht weiter verfolgt wird. Stattdessen wird der  $E_{n;i}$ -Wert wie in Gleichung (2.8) dargestellt mit  $k = 2$  verwendet.

### 5.1.1 Charakteristik der Methode

Wie sich in Gleichung (5.1) zeigt, ist der Einfluss jedes Teilnehmers auf den Referenzwert, also den arithmetischen Mittelwert, gleich, da alle Teilnehmer in Gruppe „r“ nach Gleichung (5.2) gleich gewichtet werden. Diese Tatsache ist jedoch nicht gleichbedeutend damit, dass alle Teilnehmer den gleichen Korrelationskoeffizienten  $r_{i;\text{ref};C_0}$  haben. Vielmehr ist dieser nach Gleichung (2.38) direkt proportional zur Teilnehmerunsicherheit  $u_i$ . Da der Gewichtungsfaktor des arithmetischen Mittelwerts nur von der Teilnehmerzahl in Gruppe „r“ abhängt, wird diese direkte Proportionalität nicht aufgehoben und es gilt

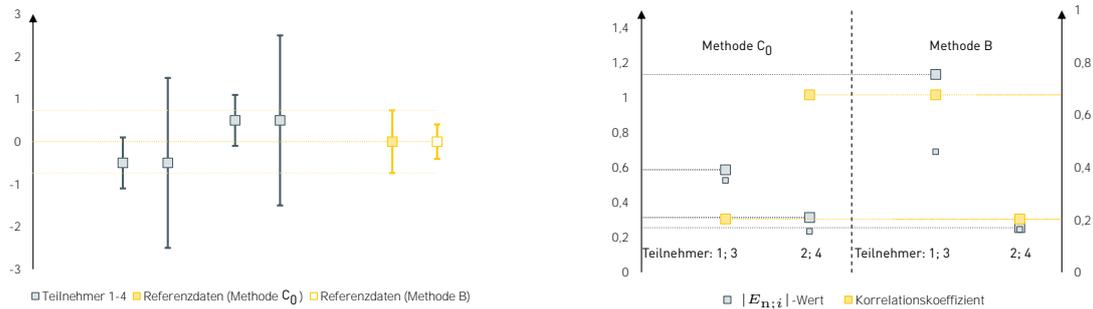
$$r_{i;\text{ref};C_0} = \frac{u_i}{N_r u_{\text{ref};C_0}} \quad (5.7)$$

für Teilnehmer in Gruppe „r“. An dieser Stelle sei angemerkt, dass erst durch die Definition von  $u_{\text{ref};C_0}$  in Gleichung (5.6) sichergestellt ist, dass der Korrelationskoeffizient  $r_{i;\text{ref};C_0}$  kleiner als eins ist. Würde Methode  $C_0$  lediglich auf die empirische Standardabweichung zurück greifen, so kann diese je nach Datenlage auch kleiner sein als eine einzelne Teilnehmerunsicherheit und ein Korrelationskoeffizient größer eins würde sich für den entsprechenden Teilnehmer ergeben. Da Korrelationskoeffizienten sinnvollerweise nur Werte zwischen 0 und 1 zugeordnet werden, zeigt sich an dieser Stelle die Notwendigkeit der Berücksichtigung von Gleichung (5.5) in der Referenzunsicherheit.



**Abbildung 5.2**

Gezeigt ist ein Beispieldatensatz bestehend aus vier Teilnehmern, deren Wert  $x_i$  betragsmäßig gleich ist. Die Standardunsicherheiten unterscheiden sich sukzessive um 0,20. Es zeigt sich in der rechten Graphik, dass der Korrelationskoeffizient direkt proportional zur Teilnehmerunsicherheit ist. Die kleinen grauen Quadrate in der rechten Abbildung zeigen informativ den  $E_{n;i}$ -Wert für den Fall, dass die Korrelation nicht berücksichtigt wird.



**Abbildung 5.3**

Gezeigt sind vier Teilnehmer, wobei Teilnehmer 1 und 2 den Wert  $x_{i=1;2} = -0,5$  und Teilnehmer 3 und 4 den Wert  $x_{i=3;4} = 0,5$  aufweisen. Aufgrund dieser Symmetrie ist der Referenzwert nach Methode B und C<sub>0</sub> gleich Null. Teilnehmer 1 und 3 weisen die gleiche Unsicherheit auf, mit  $u_{i=1;3} = 0,30$ . Teilnehmer 2 und 4 haben die Unsicherheit  $u_{i=2;4} = 1,0$ . Folglich sind der betragsmäßige  $E_{n;i}$ -Wert und der Korrelationskoeffizient jeweils für Teilnehmer 1 und 3, sowie für Teilnehmer 2 und 4 identisch. Wie rechts zu sehen ist, zeigt sich zwischen Methode B und Methode C<sub>0</sub> eine grundlegend andere Charakteristik. Während nach Methode C<sub>0</sub> Teilnehmer 1 und 3 einen kleinen Korrelationskoeffizienten aufweisen, ist dieser nach Methode B für die beiden genannten Teilnehmer aufgrund ihrer kleinen Unsicherheit groß. Umgekehrt ist es für Teilnehmer 2 und 4. Aufgrund ihrer großen Unsicherheit weisen diese beiden Teilnehmer gemäß Methode C<sub>0</sub> einen großen Korrelationskoeffizienten auf, nach Methode B allerdings einen eher kleinen. Die konkrete Auswirkung der Korrelation zeigt sich letztlich im  $E_{n;i}$ -Wert.

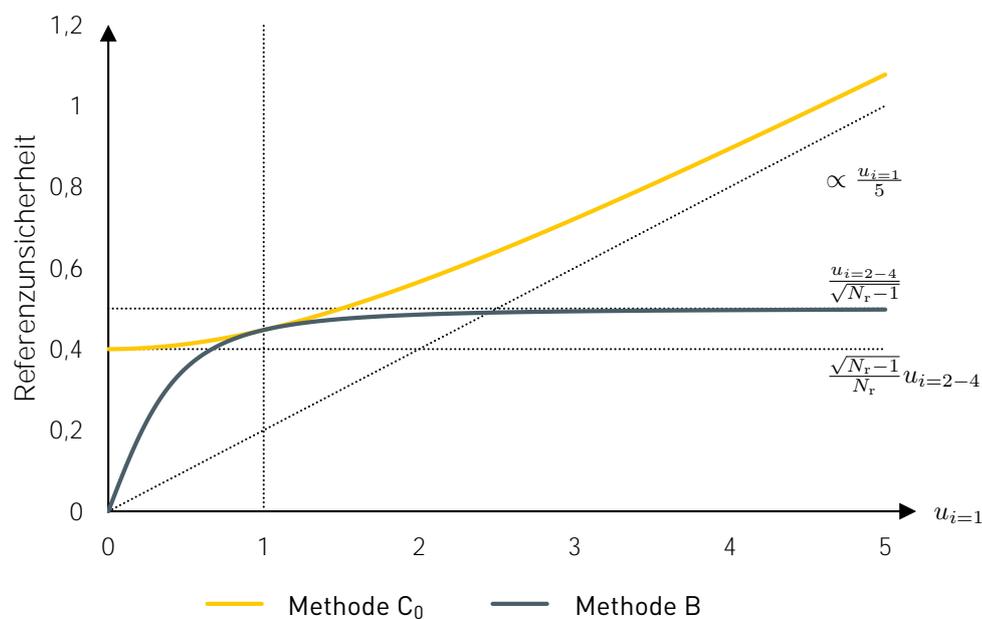
Je größer die Teilnehmerunsicherheit, desto größer ist auch der Einfluss der Korrelation in der Leistungsbewertung, wie Abbildung 5.2 zeigt. Der  $E_{n;i}$ -Wert für Gruppe „r“ ist in diesem Fall gegeben durch

$$E_{n;i} = \frac{x_i - x_{\text{ref};C_0}}{2\sqrt{(1 - \frac{2}{N_r})u_i^2 + u_{\text{ref};C_0}^2}} \quad \text{für } i \text{ in Gruppe „r“}. \quad (5.8)$$

Der  $E_{n;i}$ -Wert ist immer definiert, da  $N_r \geq 2$ , bzw. sogar  $N_r > 2$ , wie in Kapitel 8 noch gezeigt wird. Gleichungen (5.7) und (5.8) weisen einen deutlichen Unterschied zu Methode B auf. Während in Methode B nach Gleichung (4.10) vor allem Teilnehmern mit einer verhältnismäßig kleinen Unsicherheit eine große Korrelation zugeordnet wird, wird in Methode C<sub>0</sub> Teilnehmern mit einer verhältnismäßig großen Unsicherheit eine große Korrelation zugeordnet, obgleich der Einfluss auf den Referenzwert für alle Teilnehmer gleich ist. Die Auswirkung der Korrelation zeigt sich erst in der Leistungsbewertung über den  $E_{n;i}$ -Wert. Ein Vergleich dieser unterschiedlichen Charakteristiken von Methode B und C<sub>0</sub> ist in Abbildung 5.3 gezeigt.

In den bisher betrachteten Beispielen ist die Referenzunsicherheit durch die empirische Standardabweichung aus Gleichung (5.4) gegeben. Nachfolgend wird nun der Fall betrachtet, dass die Referenzunsicherheit durch die Fehlerfortpflanzung des arithmetischen Mittelwertes, d.h. durch Gleichung (5.5) gegeben ist. Um zu vermeiden, dass die Referenzunsicherheit von Methode C<sub>0</sub> unterschätzt wird, kann es sinnvoll sein, die Referenzunsicherheit gemäß Gleichung (5.6) zu definieren. Es sei beispielsweise angenommen, dass  $N_r = 5$  Teilnehmer mit gleichem Wert  $x_i = 0 \quad \forall i \in [1; 5]$  vorliegen. Die Teilnehmer 2 bis 5 berichten vereinfacht die gleichen Unsicherheiten  $u_i = 1 \quad \forall i \in [2; 5]$ , die Unsicherheit des Teilnehmers 1 hingegen sei variabel. Alle Teilnehmerwerte sind identisch, d.h. der Referenzwert ist nach Methode B und C<sub>0</sub> gegeben durch  $x_{\text{ref};B} = x_{\text{ref};C_0} = 0$ . Da keine Streuung in den Teilnehmerdaten

vorliegt, ist  $u_{\text{ref};C_0;1} = 0$ , sodass die Referenzunsicherheit nur durch  $u_{\text{ref};C_0;2}$  aus Gleichung (5.5) gegeben sein kann. Der Verlauf der Referenzunsicherheiten  $u_{\text{ref};B}$  und  $u_{\text{ref};C_0} = u_{\text{ref};C_0;2}$  ist in Abbildung 5.4 gezeigt. Signifikant ist der Unterschied, dass im gewählten Beispiel die Referenzunsicherheit nach Methode B nach oben begrenzt ist und nicht größer wird als  $u_{i=2-4}/\sqrt{N_r-1} = 1/2$ . Für eine größer werdende Unsicherheit des Teilnehmers 1 konvergiert die Referenzunsicherheit von Methode B von unten gegen diesen Grenzwert. Der Grund dafür ist, dass wir bereits in Kapitel 4 diskutiert, Methode B vor allem von kleinen Unsicherheiten dominiert wird und verhältnismäßig große Unsicherheiten keinen Einfluss auf die Referenzunsicherheit haben. Dementsprechend ist die Referenzunsicherheit nach Methode B nach unten nur durch die 0 begrenzt. Anders verhält es sich mit der Fehlerfortpflanzung aus Gleichung (5.5). Diese ist proportional zu den einzelnen Teilnehmerunsicherheiten und konvergiert somit für eine größer werdende Unsicherheit  $u_{i=1}$  von Teilnehmer 1 gegen  $u_{i=1}/5$ . Dadurch ist klar, dass die Referenzunsicherheit von Methode  $C_0$  in diesem Fall beliebig groß werden kann. Damit verbunden ist eine mögliche Überschätzung der Referenzunsicherheit, was letztlich zu einer zu guten Leistungsbewertung der Teilnehmer führen kann.



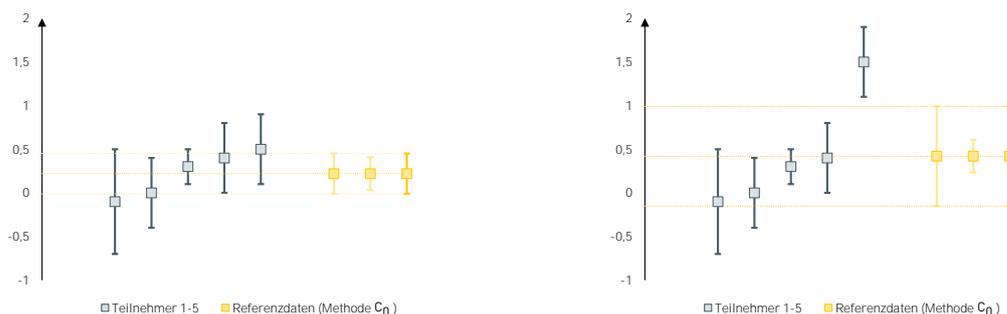
**Abbildung 5.4**

*Im Beispiel  $x_i = 0 \forall i \in [1; N_r]$  mit  $N_r = 5$  und  $u_i = 1 \forall i \in [2; N_r]$ , ist der Verlauf der Referenzunsicherheit der Methoden B und  $C_0$  in Abhängigkeit von der variablen Teilnehmerunsicherheit  $u_{i=1}$  gezeigt. Die Referenzunsicherheit ist in Methode B nach oben begrenzt, nach unten jedoch nur durch die 0. In Methode  $C_0$  kann die Referenzunsicherheit aus Gleichung [5.5] hingegen beliebig groß werden, ist dafür jedoch nach unten begrenzt.*

Je nach Datenlage ist die Wahl der Referenzunsicherheit nach Gleichung (5.6) daher problematisch und genau abzuwägen. Während  $u_{\text{ref};C_0;1}$  zu einer Unterschätzung der Referenzunsicherheit führen kann und nach unten nur durch die 0 begrenzt ist, kann  $u_{\text{ref};C_0;2}$  zu einer Überschätzung der Referenzunsicherheit führen, ist dafür aber nach unten begrenzt. Hier zeigt sich das Defizit von Methode  $C_0$ , dass Teilnehmerwert  $x_i$  und Teilnehmerunsicherheiten  $u_i$  in der Bestimmung der Referenzdaten nicht gemeinsam betrachtet werden. Eine Behebung dieses Defizits ist durch Methode D aus Kapitel 6 gegeben.

## 5.2 Robuster Algorithmus (Methode C)

Die grundlegenden Gleichungen des arithmetischen Mittelwertes (5.1) und (5.4) zeigen, dass dieser anfällig ist für Ausreißer, da alle  $N_r$  Teilnehmer aus Gruppe „r“ gleichermaßen in die Bestimmung der Referenzdaten einfließen. Ein Beispiel dafür ist in Abbildung 5.5 gegeben. Dort zeigt sich, dass Methode  $C_0$  in der in Abschnitt 5.1 beschriebenen Form nicht robust ist, d.h. leicht durch Ausreißer beeinflussbar ist. Ein Kriterium zur Eliminierung oder Modifikation von Ausreißern ist in der bisher beschriebenen Methode  $C_0$  nicht gegeben.



**Abbildung 5.5**

Die in Abschnitt 5.1 beschriebene Methode  $C_0$  ist sowohl im Referenzwert, als auch u.U. in der Referenzunsicherheit anfällig für Ausreißer. Im gezeigten Beispiel wurde zwischen den beiden Abbildungen lediglich der Wert  $x_5$  von Teilnehmer 5 geändert.

Eine Möglichkeit den arithmetischen Mittelwert hinsichtlich seiner Robustheit zu verbessern ist durch den Algorithmus aus Anhang C, Abschnitt 3.1 der ISO 13528:2022 [3] gegeben. Durch die Anwendung dieses Algorithmus wird der Einfluss von Ausreißern auf die Referenzdaten reduziert, sodass eine robuste Auswertungsmethode resultiert. Die Grundlage zu diesem Algorithmus findet sich in Ansätzen in [20] wieder. Die nachfolgenden Darstellungen werden im Rahmen dieses Dokuments als Methode C bezeichnet. Voraussetzung für die Anwendbarkeit von Methode C ist, dass mindestens vier Teilnehmerdatensätze in die Bestimmung der Referenzdaten einfließen, d.h.  $N_r \geq 4$ . Der Ausreißeranteil in den Daten darf zudem nicht größer sein als 25%. Hierfür ist ggf. eine graphische Analyse heranzuziehen. Wie später noch klar wird, geht Methode C davon aus, dass die vorliegenden Werte  $x_i$  der Teilnehmer Stichproben einer zugrundeliegenden und unbekanntem Normalverteilung sind.

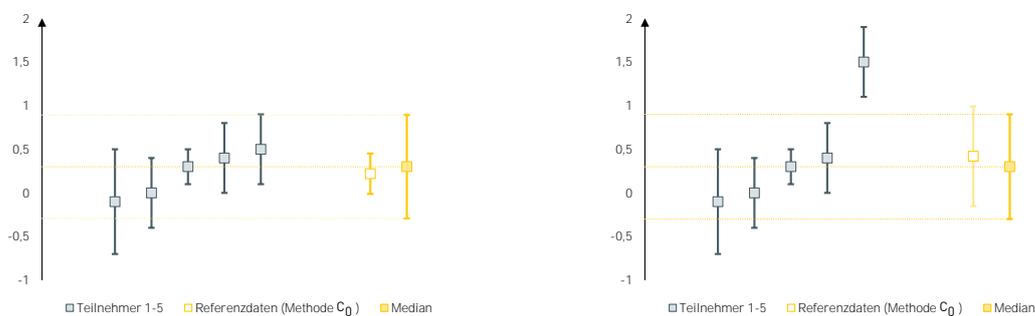
### 5.2.1 Berechnung der Referenzdaten

Zur Vereinfachung sei oBdA angenommen, dass die Teilnehmerwerte aufsteigend sortiert sind, d.h.  $x_i \leq x_{i+1} \forall i \in [1; N_r - 1]$ . Zudem wird  $\mathbf{x} \equiv \{x_1; \dots; x_{N_r}\}$  definiert. In jedem Iterationsschritt  $n$  des Algorithmus werden ein Mittelwert  $x_{C;n}$  und eine Standardabweichung  $u_{C;n}$  ermittelt. Der Anfangswert des Mittelwertes ist durch den Median („med“) gegeben, d.h.

$$x_{C;n=0} = \text{med}(\mathbf{x}) = \begin{cases} x_{i=\frac{N_r+1}{2}} & N_r \text{ ungerade,} \\ \frac{x_{i=\frac{N_r}{2}} + x_{i=\frac{N_r}{2}+1}}{2} & N_r \text{ gerade.} \end{cases} \quad (5.9)$$

Der Median ist im Vergleich zum arithmetischen Mittelwert deutlich robuster, da dieser nur von den

konkreten Werten im mittleren Bereich der sortierten Werte von  $\mathbf{x}$  abhängt. Die restlichen Werte ( $i$  nahe 1 oder  $N_r$ ) spielen dabei nur in ihrer relativen Lage zu den Werten im mittleren Bereich eine Rolle, nicht aber in ihren konkreten Zahlenwerten. In Abbildung 5.5 ist beispielsweise zu sehen, dass sich der arithmetische Mittelwert bei der Änderung eines Teilnehmerwertes signifikant ändert. Das gleiche Beispiel ist für den Median in Abbildung 5.6 gezeigt. Im Beispiel dort ist der Median durch den Wert von Teilnehmer  $i = 3$  gegeben unabhängig davon, welchen Wert Teilnehmer 5 annimmt. Wichtig ist nur, dass der Wert von Teilnehmer 5 größer bleibt als der Wert von Teilnehmer 3, d.h. dass sich die Sortierung der Werte nach ihrer Größenordnung nicht ändert. Das Beispiel zeigt, dass für den Median lediglich die Datenanzahl  $N_r$ , die relative Lage der Werte zueinander, sowie die konkreten Werte von einem bzw. maximal zwei Teilnehmer ausschlaggebend sind.



**Abbildung 5.6**

Im Beispiel von Abbildung 5.5 zeigt sich, dass der arithmetische Mittelwert nicht robust ist. Im Gegensatz dazu ändert sich der Median im gezeigten Beispiel nicht, da sich von Teilnehmer 5 nur der konkrete Werte, nicht aber die relative Lage zum Wert in der Mitte von  $\mathbf{x}$  ändert.

Der Anfangswert der Standardabweichung  $u_{C;n=0}$  ist im Wesentlichen durch die sogenannte „median absolute deviation“  $\text{MedAD}(\mathbf{x}) = \text{med}(|\mathbf{x} - x_{C;n=0}|)$  gegeben

$$u_{C;n=0} = 1,483 \cdot \text{med}(|\mathbf{x} - x_{C;n=0}|) = 1,483 \cdot \text{MedAD}(\mathbf{x}). \quad (5.10)$$

Ähnlich zur empirischen Standardabweichung wird die absolute Differenz aller Werte  $x_i \in \mathbf{x}$  zum Median betrachtet. Der Faktor 1,483 ist nötig, um den Median der absoluten Abweichungen  $\text{MedAD}(\mathbf{x})$  der Standardabweichung einer entsprechenden Normalverteilung anzugleichen, d.h. gesucht ist  $u_{C;n=0}$  mit

$$\frac{1}{\sqrt{2\pi}u_{C;n=0}} \int_{-u_{C;n=0}}^{u_{C;n=0}} e^{-\frac{x^2}{2u_{C;n=0}^2}} dx = \text{erf}\left(\frac{1}{\sqrt{2}}\right), \quad (5.11)$$

wobei  $\text{erf}(z)$  die Gaußsche Fehlerfunktion entsprechend Gleichung [2.25] ist. Per Definition liegen 50 % der Werte im Intervall  $\text{med}(\mathbf{x}) \pm \text{MedAD}(\mathbf{x})$ , d.h.

$$\frac{1}{\sqrt{2\pi}u_{C;n=0}} \int_{-\text{MedAD}(\mathbf{x})}^{\text{MedAD}(\mathbf{x})} e^{-\frac{x^2}{2u_{C;n=0}^2}} dx = \text{erf}\left(\frac{\text{MedAD}(\mathbf{x})}{\sqrt{2}u_{C;n=0}}\right) = \frac{1}{2}. \quad (5.12)$$

Daraus folgt

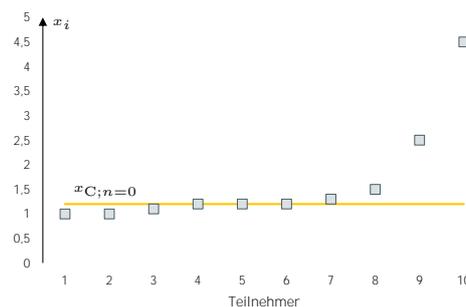
$$u_{C;n=0} = \frac{1}{\sqrt{2}\text{erf}^{-1}(1/2)} \cdot \text{MedAD}(\mathbf{x}) \simeq 1,483 \cdot \text{MedAD}(\mathbf{x}), \quad (5.13)$$

wobei  $\text{erf}^{-1}(z)$  die inverse Funktion der Gaußschen Fehlerfunktion bezeichnet.

Falls  $u_{C;n=0} = 0$  ist, kann der Algorithmus nicht loslaufen. In diesem Fall wird Gleichung (5.10) durch die ausreißerbereinigte empirische Stichprobenstandardabweichung ersetzt. Diese Ersetzung betrifft nur den Anfangswert  $u_{C;n=0}$ , der restliche Algorithmus ändert sich dadurch nicht und ist nachfolgend beschrieben. Um die nun folgenden Schritte auch beispielhaft zu veranschaulichen, wird der Datensatz

$$\begin{aligned} \mathbf{x} &= \{x_1; x_2; x_3; x_4; x_5; x_6; x_7; x_8; x_9; x_{10}\} \\ &= \{1,0; 1,0; 1,1; 1,2; 1,2; 1,2; 1,3; 1,5; 2,5; 4,5\} \end{aligned} \quad (5.14)$$

verwendet, der in Abbildung 5.7 gezeigt ist. Da die Unsicherheiten der Teilnehmer für den Algorithmus keine weitere Rolle spielen, sind diese nachfolgend nicht näher spezifiziert. Die Werte der letzten beiden Teilnehmer 9 und 10 können als Ausreißer erachtet werden, sodass der Ausreißeranteil bei 20% liegt und somit die Voraussetzungen für Methode C erfüllt sind.



#### Abbildung 5.7

Gezeigt sind die Werte der Teilnehmer aus Gleichung (5.14) und der daraus resultierende Median, der den Startwert  $x_{C;n=0}$  definiert.

Ausgehend von den Anfangswerten  $x_{C;n=0}$  und  $u_{C;n=0}$  wird in jedem Iterationsschritt  $n$  ein  $\delta_n$  definiert

$$\delta_n = l \cdot u_{C;n-1} \quad \text{mit} \quad l = \frac{3}{2}. \quad (5.15)$$

Damit wird eine  $\delta_n$ -Umgebung um den Wert  $x_{C;n-1}$  eröffnet, wie in Abbildung 5.9 gezeigt. Alle Werte, die außerhalb dieser Umgebung liegen, werden als Ausreißer erachtet und modifiziert, indem sie auf die Grenzen der  $\delta_n$ -Umgebung geschoben werden, d.h.

$$x_{i;n} = \begin{cases} x_{C;n-1} - \delta_n & x_i < x_{C;n-1} - \delta_n \\ x_{C;n-1} + \delta_n & x_i > x_{C;n-1} + \delta_n \\ x_i & \text{sonst} \end{cases} \quad (5.16)$$

Durch die Modifikation der Ausreißer entsteht ein neuer Datensatz  $\mathbf{x}_n$ . An dieser Stelle ist es wichtig darauf hinzuweisen, dass in jedem Iterationsschritt der Ausgangsdatsatz  $\mathbf{x}$  mit den Originalteilnehmerdaten  $x_i$  modifiziert wird. Der Datensatz  $\mathbf{x}_n$  mit den Werten  $x_{i;n}$  ist lediglich ein Hilfsdatensatz um die nachfolgenden Berechnungen durchzuführen. Für den Beispieldatensatz aus Gleichung (5.14) ergibt sich im ersten Iterationsschritt

$$\mathbf{x}_{n=1} = \{1,0; 1,0; 1,1; 1,2; 1,2; 1,2; 1,3; 1,5; 1,53; 1,53\}. \quad (5.17)$$

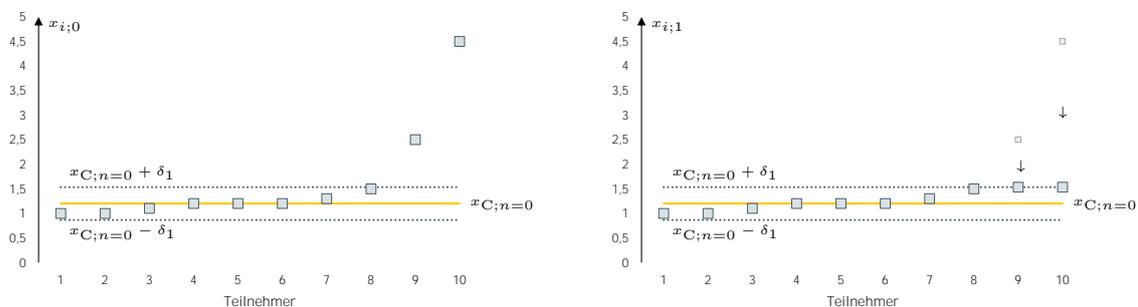
Abbildung 5.9 zeigt die Datenmodifikation nach Gleichung (5.16) für den ersten Iterationsschritt  $n = 1$  anhand des Beispieldatensatzes aus Gleichung (5.14). Mit den neuen Daten  $x_{i;n}$ , die durch Gleichung (5.16) erhalten wurden, wird ein neuer arithmetischer Mittelwert

$$x_{C;n} = \frac{1}{N_r} \sum_{i=1}^{N_r} x_{i;n} \quad (5.18)$$

sowie die entsprechende Unsicherheit über die dazugehörige empirische Standardabweichung berechnet

$$u_{C;n} = 1,134 \cdot \sqrt{\sum_{i=1}^{N_r} \frac{(x_{i;n} - x_{C;n})^2}{N_r - 1}}. \quad (5.19)$$

Der Faktor 1,134 wird am Ende dieses Abschnitts genauer betrachtet.



**Abbildung 5.8**

Ausgehend von den Startwerten  $x_{C;n=0}$  und  $u_{C;0}$  wird im ersten Iterationsschritt  $n = 1$  nach Gleichung (5.15) eine  $\delta_1$ -Umgebung definiert. Die Ausgangsdaten der Teilnehmer werden modifiziert und auf die Grenzen der  $\delta_1$ -Umgebung verschoben, falls diese sich außerhalb der  $\delta_1$ -Umgebung befinden. Dadurch entsteht ein neuer Datensatz  $x_{i;1}$ .

Die Berechnungen nach den Gleichungen (5.15), (5.16), (5.18) und (5.19) definieren einen Iterationsschritt und werden sukzessive ausgeführt und wiederholt, bis sich bezüglich  $x_{C;n}$  und  $u_{C;n}$  eine Konvergenz einstellt und sich das Ergebnis dieser Werte in der dritten signifikanten Stelle nicht mehr ändert. Abbildung 5.9 zeigt das Konvergenzverhalten für den Beispieldatensatz aus Gleichung (5.14).

Sei  $n_c$  der Iterationszyklus, bei dem diese Konvergenz festgestellt wird, so können daraus die finalen Referenzdaten für Methode C definiert werden. Für den Referenzwert  $x_{\text{ref};C}$  gilt

$$x_{\text{ref};C} = x_{C;n_c} \quad (5.20)$$

und für die dazugehörige Unsicherheit folgt

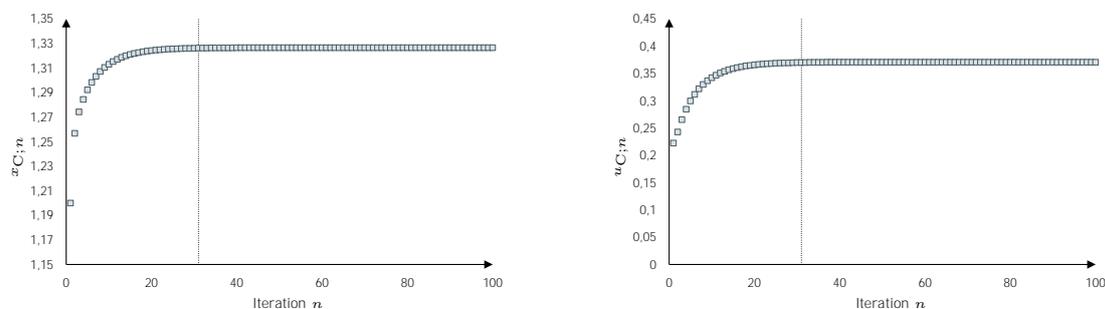
$$u_{\text{ref};C;1} = \frac{u_{C;n_c}}{\sqrt{N_r}}. \quad (5.21)$$

Hier sei angemerkt, dass Gleichung (5.21) im Vergleich zu ISO 13528:2022 [3] modifiziert wurde, da dort die Division mit dem Faktor  $\sqrt{N_r}$  fehlt. Da es sich bei der Referenzunsicherheit um die Unsicherheit eines Mittelwertes handelt, ist dieser zu ergänzen, wie in Gleichung (5.21) gezeigt. Dadurch ist klar, dass die Referenzunsicherheit mit steigender Teilnehmerzahl  $N_r$  in Gruppe „r“ kleiner wird. Für den Beispieldatensatz aus Gleichung (5.14) resultiert am Ende  $x_{\text{ref};C} = 1,33$  und  $u_{\text{ref};C;1} = 0,12$ .

Ähnlich zu Methode  $C_0$  ist es auch für Methode C möglich, in der finalen Referenzunsicherheit die Fehlerfortpflanzung aus Gleichung (5.5) zu berücksichtigen um die Rückführung zu gewährleisten, obgleich dies in ISO 13528:2022 nicht vorgeschlagen wird. Die Referenzunsicherheit für Methode C ist dann gegeben durch

$$u_{\text{ref};C} = \max(u_{\text{ref};C;1}; u_{\text{ref};C;2}). \quad (5.22)$$

Für die Anwendung von Gleichung (5.22) sei auf die Diskussion am Ende des Abschnitts 5.2.3 verwiesen. Da für die Definition von  $u_{\text{ref};C;2}$  Details aus Abschnitt 5.2.2 benötigt werden, wird an dieser Stelle auf Gleichung (5.30) verwiesen. In den nachfolgenden Ausführungen wird der Fokus zudem auf  $u_{\text{ref};C;1}$  gelegt um die Funktionsweise des Algorithmus und dessen Auswirkung auf die empirische Standardabweichung aufzuzeigen. Falls nicht anders angegeben, wird daher in den nachfolgenden Kapitel davon ausgegangen, dass die Teilnehmerunsicherheiten  $u_i$  deutlich kleiner sind als die Streuung der Teilnehmerwerte und daher  $u_{\text{ref};C} = u_{\text{ref};C;1}$  ist.



**Abbildung 5.9**

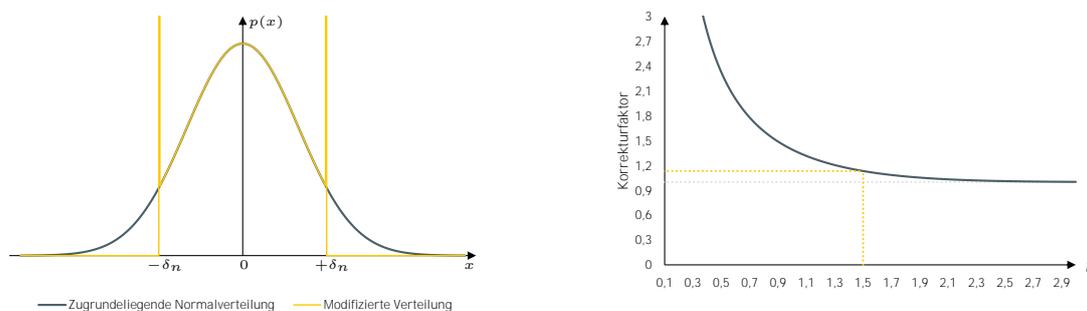
Gezeigt sind die Werte für  $x_{C;n}$  und  $u_{C;n}$  für die einzelnen Iterationen  $n$ . Die Konvergenz des Algorithmus ist zu erkennen und tritt bei  $n_c = 31$  ein.

### Zusatzinformation zum Faktor 1,134 in Gleichung (5.19)

Für den Algorithmus von Methode C ist charakteristisch, dass Ausreißer nicht eliminiert, sondern modifiziert werden und in jedem Iterationsschritt auf die Grenzen der  $\delta_n$ -Umgebung nach Gleichung (5.15) verschoben werden. Durch die Modifikation der Werte besteht die Gefahr, dass die Standardabweichung unterschätzt wird. Die Multiplikation mit dem Faktor 1,134 in Gleichung (5.19) korrigiert diesen Fehler

und wird nachfolgend hergeleitet.

Methode C geht zunächst davon aus, dass die einzelnen Werte  $x_i$  der Teilnehmer, Stichproben einer zugrunde liegenden Normalverteilung sind. In einer solchen Stichprobe sind auch Werte möglich, die von den anderen Werten stark abweichen und entsprechend der unbekannt Normalverteilung nur mit einer geringen Wahrscheinlichkeit realisiert werden. Da die zugrundeliegende Normalverteilung jedoch nicht bekannt ist, schätzt die Stichprobenstandardabweichung die Wahrscheinlichkeit der einzelnen Werte über deren Häufigkeit ab, sodass Ausreißern eine größere Wahrscheinlichkeit zugeordnet wird, als es gemäß der tatsächlichen Verteilung der Fall ist. Das trifft insbesondere dann zu, wenn die Stichprobenmenge klein ist. Der Algorithmus von Methode C korrigiert darin den Einfluss der Ausreißer durch die beschriebene Datenmodifikation, wodurch letztlich die zugrunde liegende Normalverteilung modifiziert wird, wie Abbildung 5.10 zeigt. Die modifizierte Verteilung ist schmaler als die ursprüngliche Normalverteilung und weist daher auch eine kleinere Standardabweichung auf, ein Effekt, der durch den Faktor 1,134 korrigiert wird.



**Abbildung 5.10**

Links: Durch die Datenmodifikation nach Gleichung [5.16] wird die zugrundeliegende Normalverteilung verändert, wodurch sich die resultierende Standardabweichung verringert. Der Mittelwert ist oBdA auf Null gesetzt.  
Rechts: Der Korrekturfaktor in Gleichung [5.19] hängt explizit von der Wahl der  $\delta_n$ -Umgebung ab. Je größer diese wird, d.h. desto weniger Daten modifiziert werden, desto näher ist der Korrekturfaktor bei 1.

Sei  $u$  die unbekannte Standardabweichung der zugrunde liegenden Normalverteilung  $p(x)$ , dessen Mittelwert oBdA gleich Null ist, dann gilt

$$p(x) = \frac{1}{\sqrt{2\pi}u} e^{-\frac{x^2}{2u^2}}. \quad (5.23)$$

Die durch Gleichung [5.16] modifizierte Wahrscheinlichkeitsdichte  $p_n(x)$  kann mit Gleichung [5.15] beschrieben werden durch

$$p_n(x) = \Theta(\delta_n - |x|) \cdot \frac{1}{\sqrt{2\pi}u} e^{-\frac{x^2}{2u^2}} + \frac{1}{2} (\delta(x - \delta_n) + \delta(x + \delta_n)) \cdot \left(1 - \operatorname{erf}\left(\frac{l}{\sqrt{2}}\right)\right), \quad (5.24)$$

wobei  $\Theta(x)$  die Heaviside-Stufenfunktion ist mit

$$\Theta(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases} \quad (5.25)$$

und  $\delta(x)$  die Dirac-Delta Distribution mit

$$\delta(x) = \begin{cases} 0 & x \neq 0 \\ \infty & x = 0 \end{cases} \quad \text{und} \quad \int_{-\infty}^{\infty} \delta(x) dx = 1. \quad (5.26)$$

Die Standardabweichung der modifizierten Wahrscheinlichkeitsdichte  $p_n(x)$  aus Gleichung (5.24) kann durch

$$u_n = \sqrt{\operatorname{erf}\left(\frac{l}{\sqrt{2}}\right) - \sqrt{\frac{2}{\pi}} l e^{-\frac{l^2}{2}} + l^2 \left(1 - \operatorname{erf}\left(\frac{l}{\sqrt{2}}\right)\right)} \cdot u \quad (5.27)$$

ausgedrückt werden. Für  $l = 3/2$  wie in Gleichung (5.15) ergibt sich somit

$$u \simeq 1,134 \cdot u_n \simeq \frac{1}{2} \cdot \left(1 - \operatorname{erf}\left(\frac{l}{\sqrt{2}}\right)\right) + 1, \quad (5.28)$$

womit der Korrekturfaktor in Gleichung (5.19) bewiesen ist. Durch die Herleitung ist zudem klar, dass der Korrekturfaktor unmittelbar mit der gewählten  $\delta_n$ -Umgebung zusammenhängt und der beschriebene Algorithmus in Methode C durch die Wahl von  $l$  diesbezüglich auch angepasst werden kann. Abbildung 5.10 zeigt die Abhängigkeit des Korrekturfaktors von der gewählten  $\delta_n$ -Umgebung. Durch die Wahl von  $l$  kann letztlich die Robustheit des Algorithmus hinsichtlich Ausreißern variiert werden. In diesem Dokument wird entsprechend [3]  $l = 3/2$  gewählt.

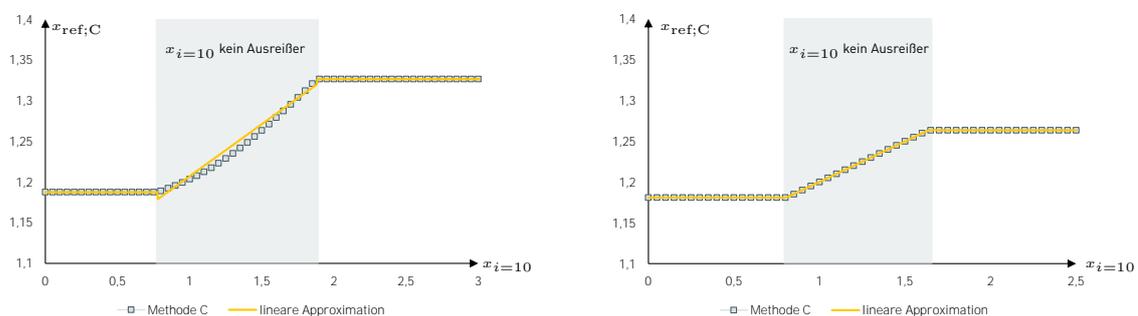
## 5.2.2 Leistungsbewertung

Der  $E_{n,i}$ -Wert, d.h. die Variable zur Leistungsbewertung ist durch Gleichung (2.8) gegeben, woraus insbesondere die Notwendigkeit zur Berücksichtigung von Korrelationen klar wird. Im linearen Modell aus Abschnitt 2.2.4 kann die Kovarianz gemäß Gleichung (2.35) in Abhängigkeit von Gewichtungsfaktor  $w_i$  und Teilnehmervarianz  $u_i^2$  ausgedrückt werden. Für den arithmetischen Mittelwert von Methode C<sub>0</sub> kann das lineare Modell direkt angewandt werden und die Kovarianz ist in Gleichung (5.3) bzw. der Korrelationskoeffizient in Gleichung (5.7) gegeben.

Da in Methode C Teilnehmerdaten in die Bestimmung der Referenzdaten einfließen, ist eine Berücksichtigung der Korrelation in der Leistungsbewertung zwingend notwendig, da andernfalls die Leistungen der Teilnehmer als zu gut bewertet werden. Allerdings ist der lineare Zusammenhang zwischen Teilnehmerwert  $x_i$  und Referenzwert  $x_{\text{ref};C}$  nicht eindeutig gegeben. Teilnehmer, deren Werte  $x_i$  durch den Algorithmus nicht modifiziert werden (d.h.  $x_i = x_{i;n_c}$ ), fließen mit einer Mindestgewichtung  $w_i = 1/N_T$  in den Referenzwert ein. Für Teilnehmer, deren Wert modifiziert wird (d.h.  $x_i \neq x_{i;n_c}$ ), verringert sich die Korrelation. Da die Art und Weise, wie diese Daten modifiziert werden, von den restlichen Daten abhängt, ist es naheliegend, dass sich die Korrelation der anderen Teilnehmer mit dem Referenzwert, erhöht. Letzteres kann evtl. numerisch abgeschätzt werden, indem der Wert des betreffenden Teilnehmers leicht variiert wird und die Auswirkung auf den Referenzwert untersucht wird. Da dies i.A. aufwändig ist, werden für Methode C näherungsweise folgende Gewichtungsfaktoren, die letztlich in die Kovarianz einfließen, angesetzt

$$w_{i;C} = \begin{cases} \frac{1}{N_r^*} & \text{Teilnehmer in Gruppe „r“ und } x_i = x_{i;n_c} \\ 0 & \text{Teilnehmer in Gruppe „r“ und } x_i \neq x_{i;n_c} \\ 0 & \text{Teilnehmer in Gruppe „n“} \end{cases}, \quad (5.29)$$

wobei  $N_r^*$  die Anzahl der Teilnehmerdaten aus Gruppe „r“ ist, die durch den Algorithmus nicht modifiziert wird. Die Gewichtungsfaktoren sind normiert und erfüllen Gleichung (2.31). Wird kein Teilnehmerwert modifiziert, dann resultiert die Gewichtung des arithmetischen Mittelwertes. Abbildung 5.11 zeigt anhand des Beispieldatensatzes aus Gleichung (5.14) die Abhängigkeit des Referenzwertes vom variablen Wert des Teilnehmers  $i = 10$ . Die Korrelation zwischen Teilnehmer- und Referenzwert kann durch die Steigung genähert werden. In Abbildung 5.11 zeigt sich, dass zwischen Referenzwert und Teilnehmerwert keine Korrelation (Steigung null) vorliegt, wenn der Teilnehmerwert modifiziert wird. Für den Bereich, in dem der Teilnehmerwert nicht modifiziert wird, liegt eine quadratische Abhängigkeit vor, die näherungsweise linearisiert werden kann. Der Grund für die quadratische Abhängigkeit liegt darin, dass der Wert von Teilnehmer 9 im Beispieldatensatz aus Gleichung (5.14) modifiziert wird und sich das auf die Korrelation zwischen Teilnehmer 10 und dem Referenzwert auswirkt. Wird der genannte Beispieldatensatz dahingehend angepasst, dass der Wert von Teilnehmer 9 auf 1,5 gesetzt wird, so resultiert das Verhalten im rechten Bild der Abbildung 5.11. Hier ist der lineare Zusammenhang exakt gegeben und entspricht dem des arithmetischen Mittelwertes, da im Datensatz keine Ausreißer vorliegen.



**Abbildung 5.11**

*Links: Für den Beispieldatensatz aus Gleichung (5.14) ist die Abhängigkeit des Referenzwertes  $x_{\text{ref};C}$  vom Wert  $x_{i=10}$  des Teilnehmers  $i = 10$  gezeigt. Wird der Teilnehmerwert durch den Algorithmus modifiziert, so ist der Referenzwert unabhängig (Steigung null) und es liegt keine Korrelation vor. Wird der Teilnehmerwert nicht modifiziert, dann liegt eine quadratische Abhängigkeit vor. Der Grund dafür ist, dass der Teilnehmerwert  $x_{i=9}$  durch den Algorithmus modifiziert wird und das Auswirkungen auf die Korrelation zwischen Teilnehmer 10 und Referenzwert hat. Der quadratische Zusammenhang kann näherungsweise linearisiert werden.*

*Rechts: Wird der Wert von Teilnehmer 9 im Beispieldatensatz aus Gleichung (5.14) auf 1,5 geändert, sodass es sich um keinen Ausreißer mehr handelt, so zeigt sich, dass in dem Bereich, in dem der Wert  $x_{i=10}$  durch den Algorithmus nicht angepasst wird, ein linearer Zusammenhang vorliegt. Der Grund dafür ist, dass durch den Algorithmus kein Wert angepasst wird und daher der lineare Zusammenhang des arithmetischen Mittelwertes resultiert.*

Wie in Methode  $C_0$  besteht auch für Methode C die Möglichkeit, die finale Referenzunsicherheit zu definieren als das Maximum zwischen der empirischen Standardabweichung, die aus dem Algorithmus resultiert und der Unsicherheit gemäß der Fehlerfortpflanzung, wie bereits in Abschnitt 5.2.1, Gleichung (5.22) dargestellt. Aufgrund der Datenmodifikation in Methode C ist allerdings auch die Fehlerfort-

pflanzung nur näherungsweise gegeben, beispielsweise unter Berücksichtigung der oben motivierten Gewichtungsfaktoren aus Gleichung (5.29). Damit lautet die Unsicherheit der Fehlerfortpflanzung von  $x_{\text{ref};C}$

$$u_{\text{ref};C;2}^2 = \frac{1}{N_r^{*2}} \sum_{i=1}^{N_r^*} u_i^2. \quad (5.30)$$

Ähnlich zu Methode  $C_0$  wird durch die Berücksichtigung der Teilnehmerunsicherheiten in der Fehlerfortpflanzung sichergestellt, dass die Referenzunsicherheit in Methode C nicht unterschätzt wird, falls die ausreißerbereinigte Streuung der Teilnehmerwerte deutlich kleiner ist als die Teilnehmerunsicherheiten. Der Fall, dass ein Ausreißerwert über eine entsprechend hohe Teilnehmerunsicherheit erklärt werden kann, wird dadurch jedoch nicht abgedeckt. Dies ist erst mit der in Kapitel 6 vorgestellten Methode D möglich. In anderen Worten bedeutet dies, dass Methode C die Konsistenz der Datenlage lediglich anhand der Teilnehmerwerte  $x_i$  beurteilt. Dem gegenüber steht Methode D, in der die Konsistenz der Daten bezüglich Teilnehmerwerte  $x_i$  und Teilnehmerunsicherheiten  $u_i$  bewertet wird, wie in Kapitel 6 noch ausführlich gezeigt wird. Für die weitere Diskussion im Rahmen dieses Dokuments wird weiter angenommen, dass die Referenzunsicherheit durch  $u_{\text{ref};C;1}$  gegeben ist. Im weiteren Verlauf dieses Kapitels wird die oben getroffene Annahme, dass  $u_{\text{ref};C} = u_{\text{ref};C;1}$  ist aufrecht erhalten, d.h. es wird angenommen, dass die Unsicherheit der Teilnehmer klein ist im Vergleich zur vorliegenden Streuung der Teilnehmerwerte.

Die Kovarianz zwischen Teilnehmer- und Referenzwert ist nach Gleichung (5.29) und (2.35) gegeben durch

$$u(x_i; x_{\text{ref};C}) = \begin{cases} \frac{u_i^2}{N_r^*} & \text{Teilnehmer in Gruppe „r“ und } x_i = x_{i;n_c} \\ 0 & \text{Teilnehmer in Gruppe „r“ und } x_i \neq x_{i;n_c} \\ 0 & \text{Teilnehmer in Gruppe „n“} \end{cases}, \quad (5.31)$$

bzw. der Korrelationskoeffizient lautet

$$r_{i;\text{ref};C} = \begin{cases} \frac{u_i}{N_r^* u_{\text{ref};C}} & \text{Teilnehmer in Gruppe „r“ und } x_i = x_{i;n_c} \\ 0 & \text{Teilnehmer in Gruppe „r“ und } x_i \neq x_{i;n_c} \\ 0 & \text{Teilnehmer in Gruppe „n“} \end{cases} \quad (5.32)$$

und entspricht in seiner Struktur dem aus Gleichung (5.7) für den arithmetischen Mittelwert. Wie bereits in Methode  $C_0$  wird auch hier erst durch die Definition der Referenzunsicherheit in Gleichung (5.22) sicher gestellt, dass  $0 \leq r_{i;\text{ref};C} \leq 1$ . Bei Vernachlässigung der Fehlerfortpflanzung aus Gleichung (5.22) könnten je nach Datenlage auch Korrelationskoeffizienten größer als eins resultieren. An dieser Stelle sei angemerkt, dass die angenommene Korrelation für Methode C eine Abschätzung ist. Ausgehend von der Gewichtung des arithmetischen Mittelwertes, wird die Korrelation in Methode C dahingehend angepasst, dass die Gewichtung von Teilnehmern, deren Wert modifiziert wird und deren Korrelation dann auf Null gesetzt wird, gleichmäßig auf alle anderen Teilnehmer verteilt wird. Diese Abschätzung liefert in den meisten Fällen die richtige Größenordnung der Korrelation, die insbesondere größer ist

als die des arithmetischen Mittelwertes. Je nach Datenlage kann die tatsächliche Korrelation eines Teilnehmerwertes mit dem Referenzwert jedoch auch minimal anders ausfallen und insbesondere auch etwas größer sein. Zudem sei angemerkt, dass Werte, die als Ausreißer identifiziert werden auf den Referenzwert durch ihre relative Lage, unterhalb oder oberhalb der restlichen Daten, Einfluss auf die Referenzdaten haben. Vereinfacht wird diese Korrelation auf Null gesetzt.

Ein wesentlicher Bestandteil von Methode C ist die Voraussetzung, dass maximal ein Ausreißeranteil von 25% erlaubt ist, d.h. maximal 25% der Daten werden durch den Algorithmus modifiziert, sodass

$$\frac{3}{4}N_r \leq N_r^* \leq N_r. \quad (5.33)$$

Darüber hinaus ist in Methode C vorausgesetzt, dass mindestens vier Datensätze vorliegen, d.h.  $N_r^* \geq 3$ , woraus folgt, dass

$$w_{i;C} \leq \frac{1}{3}. \quad (5.34)$$

Gemäß Gleichung [2.36] folgt daher, dass die Unsicherheit  $U_{d_i}$  im Nenner des  $E_{n;i}$ -Wertes definiert ist und größer als Null ist

$$\begin{aligned} U_{d_i} &= 2\sqrt{(1 - 2w_i)u_i^2 + u_{\text{ref}}^2} \\ &\geq 2\sqrt{\left(1 - \frac{2}{3}\right)u_i^2 + u_{\text{ref}}^2} \\ &> 0. \end{aligned} \quad (5.35)$$

Der  $E_{n;i}$ -Wert ist damit stets definiert, entsprechend

$$E_{n;i} = \begin{cases} \frac{x_i - x_{\text{ref};C}}{2\sqrt{\left(1 - \frac{2}{N_r^*}\right)u_i^2 + u_{\text{ref};C}^2}} & i \text{ Teilnehmer in Gruppe „r“ und } x_i = x_{i;n_c} \\ \frac{x_i - x_{\text{ref};C}}{2\sqrt{u_i^2 + u_{\text{ref};C}^2}} & \text{sonst} \end{cases} \quad (5.36)$$

und eine Leistungsbewertung ist für alle Teilnehmer möglich.

### 5.2.3 Charakteristik der Methode

Anhand des Beispieldatensatzes aus Gleichung [5.14] wurde Methode C in den Abschnitten 5.2.1 und 5.2.2 vorgestellt, wodurch bereits wesentliche Eigenschaften des Algorithmus aufgezeigt wurden. An dieser Stelle wird nun auf die Unterschiede und Gemeinsamkeiten mit Methode  $C_0$ , also dem arithmetischen Mittelwert aus Abschnitt 5.1, und Methode C eingegangen. Dafür wird zunächst noch einmal das Beispiel aus Abbildung 5.11 aufgegriffen und um den arithmetischen Mittelwert ergänzt, wie in Abbildung 5.12 gezeigt. Dort ist für den Datensatz

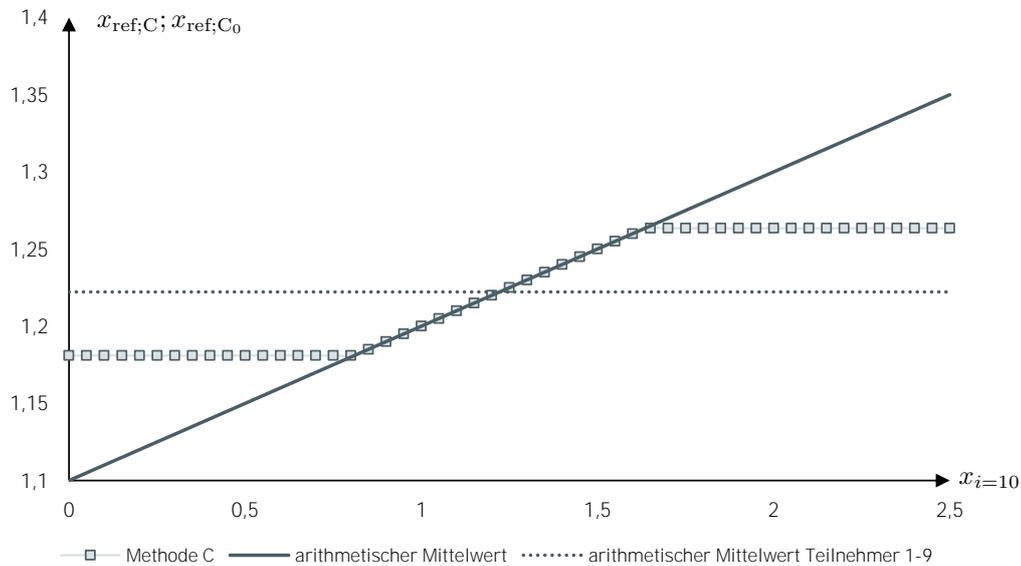
$$\begin{aligned} \mathbf{x} &= \{x_1; x_2; x_3; x_4; x_5; x_6; x_7; x_8; x_9; x_{10}\} \\ &= \{1,0; 1,0; 1,1; 1,2; 1,2; 1,2; 1,3; 1,5; 1,5; x_{10}\} \end{aligned} \quad (5.37)$$

die Abhängigkeit des Referenzwertes  $x_{\text{ref};C}$  vom Teilnehmerwert  $x_{10}$  gezeigt. Die Daten sind so gewählt, dass durch den Algorithmus die Werte  $x_1$  bis  $x_9$  nicht angepasst werden. Es zeigt sich dass Methode C mit dem arithmetischen Mittelwert übereinstimmt, falls der Wert  $x_{10}$  durch den Algorithmus nicht modifiziert wird. In den Bereichen, in denen der Algorithmus den Wert  $x_{10}$  als Ausreißer erachtet, zeigt sich der Unterschied zwischen Methode C und dem arithmetischen Mittelwert aus Methode  $C_0$ . Während sich der Einfluss des Wertes  $x_{10}$  weiterhin mit der Gewichtung  $1/10$  im arithmetischen Mittelwert niederschlägt, wird der Einfluss des Ausreißers durch den Algorithmus kompensiert. Wie sich im Vergleich mit dem arithmetischen Mittelwert der Werte  $x_1$  bis  $x_9$  jedoch zeigt, wird der Einfluss des Wertes  $x_{10}$  nicht vollständig eliminiert. Wie Abbildung 5.12 zeigt, ist der Referenzwert nach Methode C kleiner als der arithmetische Mittelwert aus den Werten  $x_1$  bis  $x_9$ , wenn es sich bei  $x_{10}$  um einen Ausreißer nach unten handelt und größer als der genannte arithmetische Mittelwert, wenn es sich um einen Ausreißer nach oben handelt. Hier wird noch einmal deutlich, dass der Algorithmus in Methode C Ausreißer nicht eliminiert, sondern modifiziert. Ein Ausreißer der kleiner ist als der Mittelwert der restlichen Teilnehmer wird durch den Algorithmus auf die untere Grenze der  $\delta_n$ -Umgebung verschoben und verringert dadurch den finalen Referenzwert. Ein Ausreißer, der größer ist als der Mittelwert der restlichen Teilnehmer wird hingegen auf die obere Grenze der  $\delta_n$ -Umgebung geschoben und erhöht damit den finalen Referenzwert. Damit wird noch einmal deutlich, dass es sich bei der Annahme für die Korrelation in Methode C in Gleichung (5.32) um eine Abschätzung handelt. Die tatsächliche Korrelation eines Wertes mit dem Referenzwert schwimmt, da durch den Algorithmus die gesamte Datenlage, d.h. die konkreten Werte der anderen Teilnehmer einen Einfluss auf die Korrelation eines Teilnehmerwertes mit dem Referenzwert hat.

Wie Abbildung 5.12 zeigt kann Methode C für das Beispiel aus Gleichung (5.37) als eine Art Kompromiss zwischen der vollen Berücksichtigung von Ausreißern im arithmetischen Mittelwert und der vollständigen Eliminierung von Ausreißern im arithmetischen Mittelwert erachtet werden.

Nun wird der Datensatz aus Gleichung (5.37) dahingehend verändert, dass neben Teilnehmerwert  $x_{10}$  ein weiterer Ausreißer hinzugefügt wird. Im Fall a, wird für Wert  $x_9$  ein Wert angenommen, der kleiner ist als der Mittelwert der Werte  $x_1$  bis  $x_8$ , d.h. es handelt sich um einen Ausreißer nach unten. Im Fall b wird für den Wert  $x_9$  ein Wert eingesetzt, bei dem es sich um einen Ausreißer nach oben handelt, d.h. der größer ist als der Mittelwert der restlichen Werten. Die Datensätze lauten damit (die Sortierung der Daten in  $\mathbf{x}$  sei für Fall a an dieser Stelle vernachlässigt)

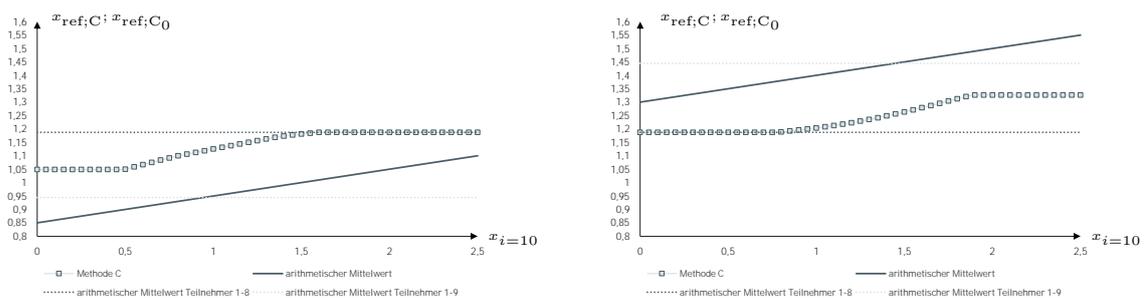
$$\begin{aligned} \mathbf{x}_a &= \{x_1; x_2; x_3; x_4; x_5; x_6; x_7; x_8; x_9; x_{10}\} \\ &= \{1,0; 1,0; 1,1; 1,2; 1,2; 1,2; 1,3; 1,5; -1,0; x_{10}\} \end{aligned} \quad (5.38)$$



**Abbildung 5.12**

Für den Datensatz (5.37) ist die Abhängigkeit des Referenzwertes  $x_{ref;C}$  aus Methode C und des Referenzwertes  $x_{ref;C_0}$  aus Methode  $C_0$  vom Teilnehmerwert  $x_{i=10}$  gezeigt. Methode C erweist sich als Kompromis zwischen der vollen Berücksichtigung und der Eliminierung von Ausreißern.

$$\begin{aligned}
 \mathbf{x}_b &= \{x_1; x_2; x_3; x_4; x_5; x_6; x_7; x_8; x_9; x_{10}\} \\
 &= \{1,0; 1,0; 1,1; 1,2; 1,2; 1,2; 1,3; 1,5; 3,5; x_{10}\}
 \end{aligned}
 \tag{5.39}$$



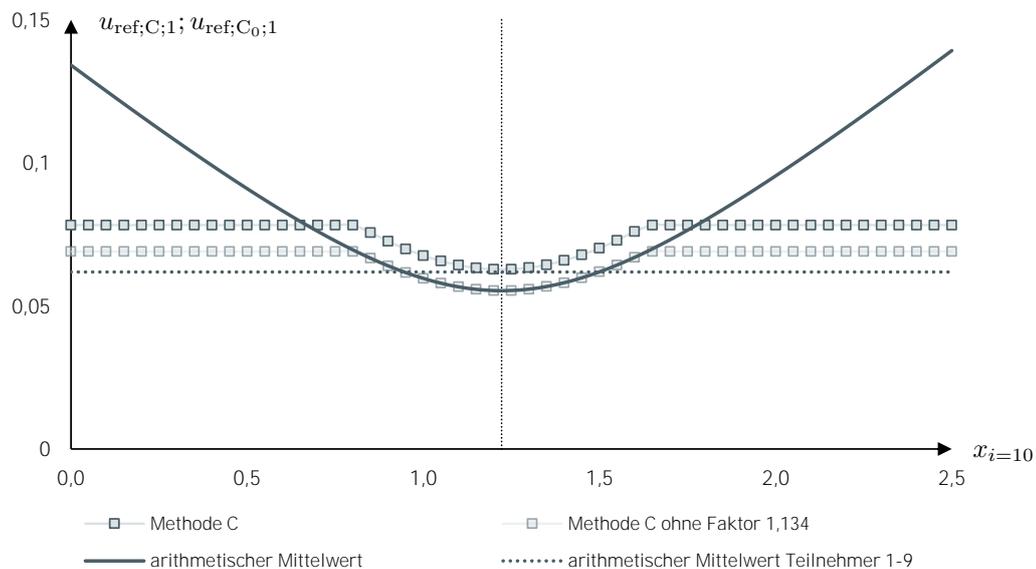
**Abbildung 5.13**

Links: Gezeigt ist der Datensatz für Fall a aus Gleichung (5.38), bei dem ein fester Ausreißer nach unten vorliegt. In dem Bereich, in dem Methode C den Wert  $x_{i=10}$  als einen Ausreißer nach oben identifiziert, heben sich beide Ausreißer auf und der arithmetische Mittelwert der Werte  $x_1$  bis  $x_8$  resultiert.

Rechts: Gezeigt ist der Datensatz für Fall b aus Gleichung (5.39), bei dem ein fester Ausreißer nach oben vorliegt. In dem Bereich, in dem Methode C den Wert  $x_{i=10}$  als einen Ausreißer nach unten identifiziert, heben sich beide Ausreißer auf und der arithmetische Mittelwert der Werte  $x_1$  bis  $x_8$  resultiert.

In Abbildung 5.13 zeigt sich, dass im Fall von mehreren Ausreißern der Referenzwert von Methode C sich deutlich von dem aus Methode  $C_0$  unterscheidet. In Fall a, bei dem ein fester Ausreißer nach unten

vorliegt, korrigiert Methode C dessen Einfluss, sodass die Referenzwerte  $x_{\text{ref};C}$  in Abhängigkeit vom Wert  $x_{10}$  im gezeigten Bereich oberhalb des arithmetischen Mittelwertes von Methode  $C_0$  liegen. Das umgekehrte Verhalten ist für Fall b zu erkennen, da dort ein fester Ausreißer nach oben vorliegt. Zudem zeigt Abbildung 5.13, dass im Fall a, wenn mit Teilnehmer 9 ein Ausreißer nach unten vorliegt, der Referenzwert nach Methode C dem arithmetischen Mittelwert der Teilnehmerwerte  $x_1$  bis  $x_8$  entspricht, wenn der Wert  $x_{10}$  vom Algorithmus als ein Ausreißer nach oben erachtet wird. Im Fall b hingegen entspricht der Referenzwert nach Methode C dem arithmetischen Mittelwert der Teilnehmer 1 bis 8, wenn der Wert  $x_{10}$  ein Ausreißer nach unten ist. Beide Fälle reproduzieren demnach den Grenzfall des arithmetischen Mittelwertes zwischen den Teilnehmerwerten  $x_1$  bis  $x_8$ , wenn im gesamten Datensatz ein Ausreißer nach oben und ein Ausreißer nach unten vorliegt. Der Grund dafür ist, dass beide Ausreißer auf die Grenzen der  $\delta_n$ -Umgebung verschoben werden. Der Ausreißer nach unten auf die untere  $\delta_n$ -Umgebungsgrenze und der Ausreißer nach oben auf die obere  $\delta_n$ -Umgebungsgrenze. Damit kompensieren sich die beiden modifizierten Ausreißer effektiv, sodass der arithmetische Mittelwert der Teilnehmer 1 bis 8 übrig bleibt. In den Bereichen, in denen effektiv zwei Ausreißer nach unten oder nach oben vorliegen, ist der Referenzwert  $x_{\text{ref};C}$  nach Methode C kleiner (Fall a) bzw. größer (Fall b) als der Mittelwert der Teilnehmer 1 bis 8, da dann beide Ausreißer auf die gleiche Grenze der  $\delta_n$ -Umgebung verschoben werden und damit ein zusätzliches Gewicht nach unten bzw. nach oben darstellen.



**Abbildung 5.14**

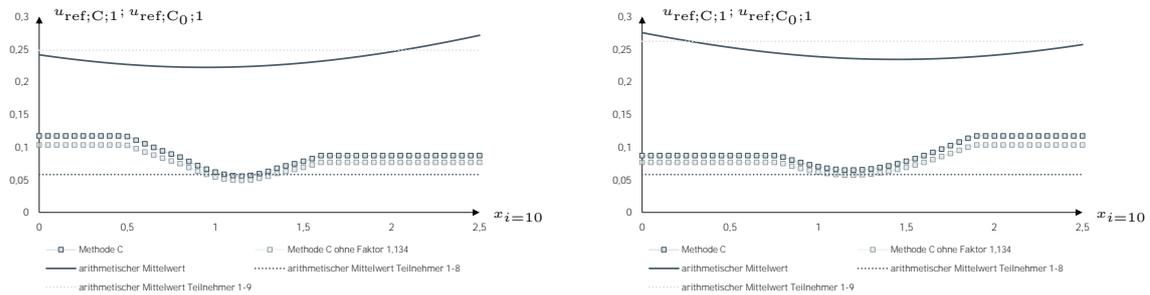
Für den Datensatz [5.37] ist die Abhängigkeit der Referenzunsicherheit  $u_{\text{ref};C;1}$  aus Methode C und der Referenzunsicherheit  $u_{\text{ref};C_0;1}$  aus Methode  $C_0$  vom Teilnehmerwert  $x_{i=10}$  gezeigt.

Bisher wurde das Verhalten des Referenzwertes  $x_{\text{ref};C}$  genauer untersucht, nun folgt die Analyse der Referenzunsicherheit  $u_{\text{ref};C;1}$ . Hierfür wird zunächst wieder der Datensatz aus Gleichung (5.37) betrachtet. Die Abhängigkeit der Referenzunsicherheit vom Teilnehmerwert  $x_{10}$  ist in Abbildung 5.14 gezeigt. Analog zum Referenzwert zeigt sich auch in der Referenzunsicherheit, dass diese in Methode C unabhängig

vom Wert des Teilnehmers 10 ist, falls dieser als Ausreißer erkannt und modifiziert wird. Hingegen ist der Einfluss eines Ausreißers in der empirischen Standardabweichung des arithmetischen Mittelwertes dominierend und die Unsicherheit  $u_{\text{ref};C_0;1}$  verhält sich näherungsweise linear bei dominierenden Ausreißerwerten. In dem Bereich, in dem der Wert  $x_{10}$  durch den Algorithmus nicht modifiziert wird, entspricht das qualitative Verhalten der Referenzunsicherheit  $u_{\text{ref};C;1}$  der empirischen Standardabweichung des arithmetischen Mittelwertes  $u_{\text{ref};C_0;1}$  aus Gleichung (5.4). Allerdings ist die Referenzunsicherheit  $u_{\text{ref};C;1}$  um einen konstanten Wert nach oben verschoben im Vergleich zu  $u_{\text{ref};C_0;1}$ . Der Grund dafür liegt im zusätzlichen Faktor 1,134, der in Gleichung (5.19) mit der empirischen Standardabweichung multipliziert wird. Eine Möglichkeit zur Anpassung von Methode C besteht darin den Korrekturfaktor in der Referenzunsicherheit auf eins zu setzen, falls keine Teilnehmerwerte durch den Algorithmus modifiziert werden, um den Grenzfall des arithmetischen Mittelwertes auch in der Unsicherheit exakt zu reproduzieren. Nachfolgend wird Methode C jedoch in der oben beschriebenen Weise verwendet, d.h. mit dem Korrekturfaktor 1,134. Falls keine Ausreißer vorliegen kann auch auf Methode  $C_0$  selbst zurück gegriffen werden. Für den Fall, dass die Werte  $x_{10}$  durch den Algorithmus nicht modifiziert werden, hängt die Referenzunsicherheit quadratisch von  $x_{10}$  ab und ist symmetrisch um den arithmetischen Mittelwert der Teilnehmer 1 bis 9, bei dem das Minimum angenommen wird. Dies liegt daran, dass an diesem Punkt der Wert  $x_{10}$  gerade dem Referenzwert entspricht, der durch den arithmetischen Mittelwert der Teilnehmer 1 bis 9 gegeben ist, sodass der Beitrag des Wertes  $x_{10}$  in der empirischen Standardabweichung verschwindet. Der Wert des Minimums liegt in der Größenordnung der empirischen Standardunsicherheit der Teilnehmer 1 bis 9. Das Minimum ohne Korrekturfaktor entspricht der empirischen Standardabweichung der Teilnehmer 1 bis 9, multipliziert mit dem Faktor  $\sqrt{(N_T - 2)/N_T}$ . Im Datenbeispiel von Gleichung (5.37) beträgt dieser Faktor ca. 1,118, weshalb das Minimum der Referenzunsicherheit von Methode C inkl. Korrekturfaktor 1,134 in Abbildung 5.14 näherungsweise der empirischen Standardunsicherheit der Teilnehmer 1 bis 9 entspricht. Je nach Teilnehmerzahl kann ist das jedoch nicht immer der Fall.

Wird nun der Datensatz um einen festen Ausreißer erweitert, wie in Gleichung (5.38) (Fall a, Ausreißer nach unten) bzw. (5.39) (Fall b, Ausreißer nach oben) beispielhaft angegeben, so resultiert das in Abbildung 5.15 und Abbildung 5.16 gezeigte Verhalten.

Wenn der Wert  $x_{10}$  qualitativ dem bereits vorliegenden Ausreißer von Teilnehmer 9 entspricht (Fall a:  $x_{10}$  Ausreißer nach unten; Fall b:  $x_{10}$  Ausreißer nach oben), ist die Referenzunsicherheit nach Methode C zum einen konstant bzgl.  $x_{10}$  und zum anderen größer als wenn es sich bei  $x_{10}$  um einen zu Teilnehmer 9 entgegengesetzten Ausreißer handelt (Fall a:  $x_{10}$  Ausreißer nach oben; Fall b:  $x_{10}$  Ausreißer nach unten). Dies liegt daran, dass zwei qualitativ entgegengesetzte Ausreißer einander im Referenzwert kompensieren, in dem Sinne, dass der eine Ausreißer auf die obere und der andere auf die untere Grenze der  $\delta_n$ -Umgebung verschoben wird. Dies wirkt sich in der gezeigten Weise auch auf die Referenzunsicherheit aus, da die Summe der mittleren quadratischen Abweichungen kleiner ist als dann, wenn beide Ausreißer auf den gleichen Wert geändert werden. Weiter fällt auf, dass die beiden konstanten Werte jeweils im entgegengesetzten Bereich identisch sind. Der Grund dafür ist, dass in einem Datensatz lediglich die Anzahl der Ausreißer und ihre qualitative Lage zum Mittelwert (d.h. Ausreißer nach oben oder nach unten) entscheidend ist. Die anderen Parameter des Algorithmus werden von den restlichen



**Abbildung 5.15**

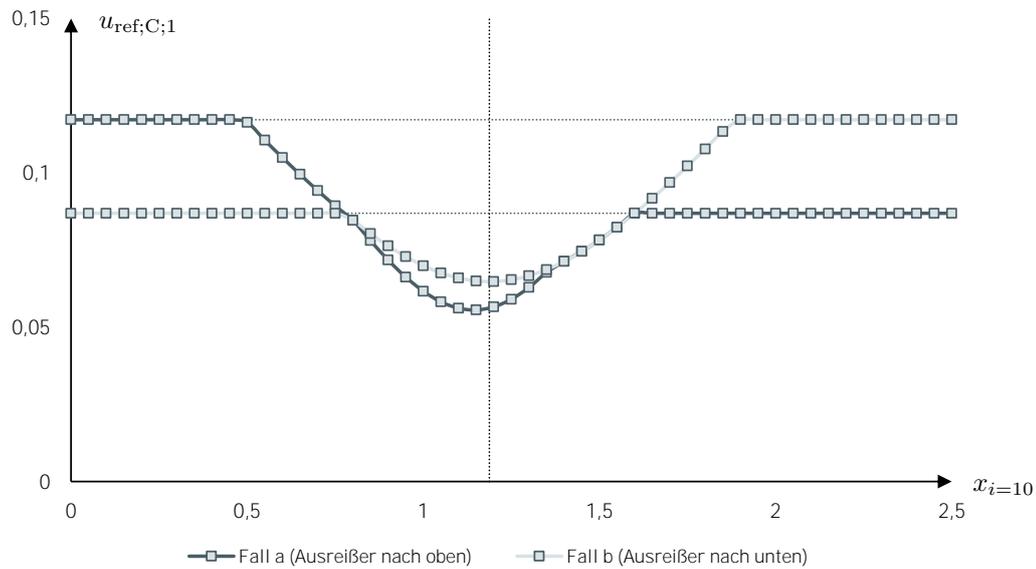
Links ist der Datensatz für Fall a aus Gleichung (5.38) gezeigt, bei dem ein fester Ausreißer nach unten vorliegt und rechts der Datensatz für Fall b aus Gleichung (5.39), bei dem ein fester Ausreißer nach oben vorliegt. In den jeweils entgegengesetzten Bereichen, in denen der Wert  $x_{i=10}$  durch den Algorithmus geändert wird, resultieren die gleichen konstanten Werte.

Teilnehmern bestimmt.

Ein leicht unterschiedliches Verhalten zeigt sich in dem Bereich, in dem der Wert von Teilnehmer  $x_{10}$  durch den Algorithmus nicht geändert wird, was insbesondere Abbildung 5.16 veranschaulicht. Das Minimum in Fall a ist kleiner als jenes von Fall b und liegt unterhalb der empirischen Standardabweichung des arithmetischen Mittelwertes der Teilnehmer 1 bis 8, wohingegen die Referenzunsicherheit von Fall b immer oberhalb dieses Wertes liegt. Grund dafür ist der in Abbildung 5.13 gezeigte Unterschied im Referenzwert, der auch für die Berechnung der Referenzunsicherheit in Methode C relevant ist. Abgesehen vom Verhalten an den beiden Minima, zeigen die Verläufe von Fall a und Fall b ein insgesamt antisymmetrisches Verhalten um den Wert  $x_{10}$ , der dem Mittelwert der Teilnehmer 1 bis 8 entspricht. Das unterschiedliche Verhalten nahe des Minimums ist auf die Datenlage der Teilnehmer 1 bis 8 zurückzuführen, da diese nicht symmetrisch verteilt um deren Mittelwert liegen. Wenn das der Fall ist, dann ist das komplette Verhalten von Fall a und Fall b über den ganzen Bereich antisymmetrisch um den Mittelwert der Teilnehmer 1 bis 8.

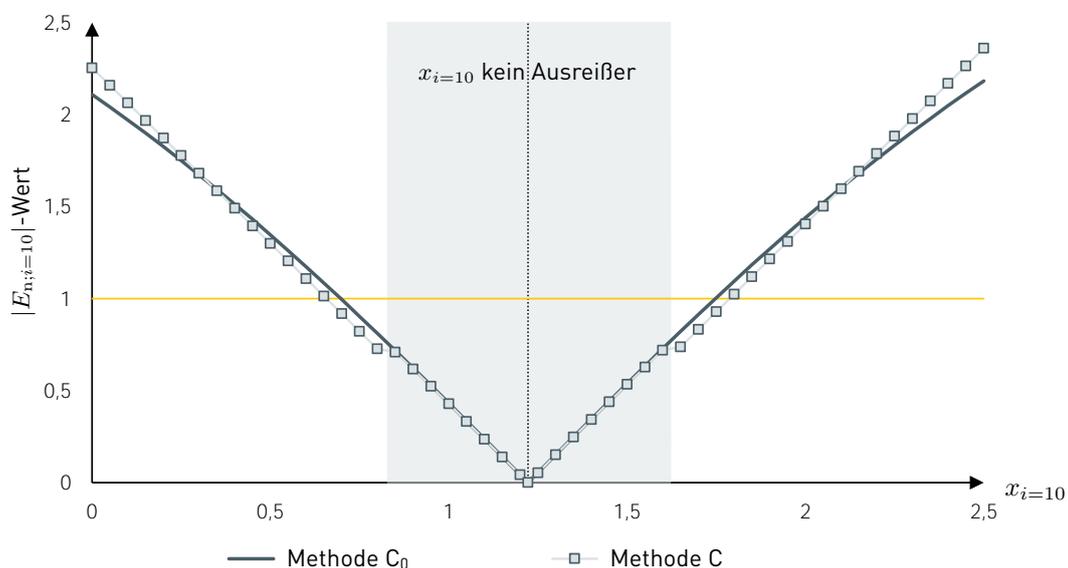
Die bisherigen Analysen des Referenzwertes und der Referenzunsicherheit in Methode C haben lediglich auf die konkreten Teilnehmerwerte  $x_i$  zurück gegriffen, da angenommen wurde, dass die Teilnehmerunsicherheiten kleiner sind als die Streuung der Werte, sodass  $u_{\text{ref};C} = u_{\text{ref};C;1}$ . Die Teilnehmerunsicherheiten werden in diesem Fall ähnlich zu Methode  $C_0$  nicht weiter berücksichtigt, sondern sind erst in der konkreten Leistungsbewertung, sowie für die Bestimmung der Kovarianz nach Gleichung (5.31) relevant.

Das Verhalten der Kovarianz bzw. des Korrelationskoeffizienten in Methode C ist ähnlich zu dem in Methode  $C_0$ , welches bereits in Abschnitt 5.1.1 ausführlich beschrieben ist. Ein Vergleich, der Leistungsbewertung zwischen Methode C und Methode  $C_0$  für den Datensatz aus Gleichung (5.37) ohne festen Ausreißer ist in Abbildung 5.17 gezeigt. Das Verhalten der Leistungsbewertung im Betrag des  $E_{n,i=10}$ -Wertes ist symmetrisch um den Mittelwert der Teilnehmer 1 bis 9. In dem Bereich, in dem der Wert  $x_{10}$  durch den Algorithmus nicht modifiziert wird, entspricht die Leistungsbewertung von Methode C der nach Methode  $C_0$ , falls für letztere die empirische Standardabweichung aus Gleichung (5.4) als



**Abbildung 5.16**

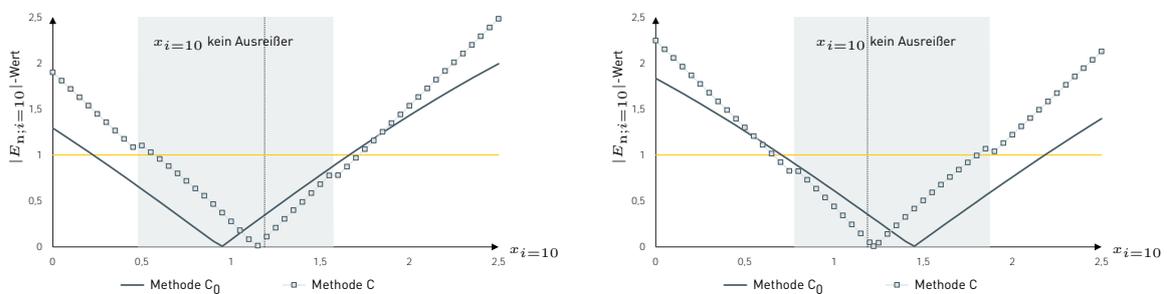
Für die Datensätze (5.38) (Fall a, ein fester Ausreißer nach unten) und (5.39) (Fall b, ein fester Ausreißer nach oben), ist die Abhängigkeit der Referenzunsicherheit  $u_{\text{ref};C;1}$  aus Methode C gezeigt. Es zeigt sich, dass die konstanten Werte der Referenzunsicherheiten in den jeweils entgegengesetzten Bereichen, in denen  $x_{10}$  nicht modifiziert wird, gleich sind und ein nahezu linearer Übergangsbereich existiert, in dem beide Fälle ähnliche Ergebnisse liefern. Die Asymmetrie im Bereich um den Mittelwert der Teilnehmer 1 bis 8 ist auf die inhomogene Datenlage zurückzuführen, da sich im gewählten Beispiel die einzelnen Werte nicht symmetrisch um den Mittelwert verteilen.



**Abbildung 5.17**

Für den Datensatz aus Gleichung (5.37) ist die Leistungsbewertung von Methode C im Vergleich zu Methode  $C_0$  gezeigt, wobei für letztere als Referenzunsicherheit Gleichung (5.4) angesetzt ist. Die Unsicherheit  $u_{10}$  wurde auf 0,25 gesetzt. Für den Fall, dass  $x_{i=10}$  nicht angepasst wird, gibt es keinen signifikanten Unterschied in der Leistungsbewertung der beiden Methoden.

Referenzunsicherheit angesetzt wird. Kleine Abweichungen können sich je nach Teilnehmerunsicherheit durch den Korrekturfaktor 1,134 aus Gleichung (5.19) ergeben, der zu einer minimal besseren Leistungsbewertung nach Methode C führen kann. Sobald der Wert von Teilnehmer 10 durch den Algorithmus als Ausreißer erkannt und modifiziert wird, ist die Leistungsbewertung nach Methode C zunächst etwas besser als nach Methode  $C_0$ , da in Methode C die Korrelation für den Teilnehmer auf Null gesetzt wird. Je weiter sich der Wert  $x_{10}$  jedoch von den anderen Teilnehmern entfernt, desto größer wird auch die Abweichung zum Referenzwert, auf den der Wert so gut wie keinen Einfluss mehr hat, abgesehen von der qualitativen Lage. Daher fällt die Leistungsbewertung nach Methode C ab einem gewissen Wert trotz nicht-vorhandener Korrelation schlechter aus als nach Methode  $C_0$ , in der der Ausreißer stets Einfluss auf den Referenzwert und dessen Unsicherheit nimmt. Insbesondere letztere ist in Methode C bei Vorliegen eines dominanten Ausreißers kleiner als nach Methode  $C_0$ , was ebenfalls zu einer schärferen Leistungsbewertung führt.



**Abbildung 5.18**

Links ist für Fall a, d.h. es existiert mindestens ein Ausreißer nach oben, die Leistungsbewertung von Methode C im Vergleich zu Methode  $C_0$  gezeigt, rechts ist diese für den Fall b, bei dem mindestens ein Ausreißer nach oben existiert gezeigt. Ein Knick in der Leistungsbewertung nach Methode C ist jeweils dann zu sehen, wenn der Wert  $x_{i=10}$  durch den Algorithmus geändert wird und die Korrelation in der Leistungsbewertung daher auf Null gesetzt wird. Die Leistungsbewertung nach Methode  $C_0$  ist verschoben, da für diese entsprechend Abbildung 5.13 der Referenzwert bereits verschoben ist. Die Teilnehmerunsicherheit  $u_{10}$  wurde auf 0,25 gesetzt. Für die Referenzunsicherheit von Methode  $C_0$  wird Gleichung (5.4) angesetzt.

Für die Datensätze aus Gleichung (5.38) (Fall a, ein fester Ausreißer nach unten) und Gleichung (5.39) (Fall b, ein fester Ausreißer nach oben), ist das Verhalten der Leistungsbewertung in Abbildung 5.18 gezeigt. Die Leistungsbewertungen von Fall a und Fall b sind zueinander antisymmetrisch um den Mittelwert der Teilnehmer 1 bis 8. Die Asymmetrie der Referenzunsicherheit aus Abbildung 5.16 ist in der Leistungsbewertung nicht zu erkennen, da die Referenzunsicherheit im Vergleich zur gewählten Teilnehmerunsicherheit von  $u_{10} = 0,25$  nicht signifikant ist. Sobald der Wert  $x_{10}$  durch den Algorithmus modifiziert wird, entsteht ein kleiner Knick in der Leistungsbewertung, da die Korrelation sprunghaft auf Null gesetzt wird, sodass die Leistung des Teilnehmers 10 letztlich besser bewertet wird. Die Leistungsbewertung nach Methode  $C_0$  ist im Vergleich zu Methode C versetzt, was vor allem an den unterschiedlichen Referenzwerten liegt, wie bereits in Abbildung 5.13 gezeigt. Zudem fällt die Leistungsbewertung nach Methode  $C_0$  tendenziell besser aus, was vor allem an der größeren Referenzunsicherheit liegt, wie in Abbildung 5.15 zu sehen ist.

Insgesamt zeigt sich, dass die Leistungsbewertung bzgl. Methode C qualitativ vergleichbar ist mit der Leistungsbewertung nach Methode C<sub>0</sub>. Unterschiede ergeben sich jedoch quantitativ, sobald Ausreißer vom Algorithmus erkannt und modifiziert werden. Dies wirkt sich sowohl auf den Referenzwert und dessen Unsicherheit, als auch auf die Annahme der Korrelation und letztlich die Leistungsbewertung aus. Wie in den Abbildungen 5.17 und 5.18 gezeigt, resultiert für einen Wert am Übergangspunkt zwischen „der Wert  $x_{10}$  wird als Ausreißer erachtet“ und „der Wert  $x_{10}$  wird nicht als Ausreißer erachtet“ eine Änderung (Knick bzw. Sprung) in der Leistungsbewertung. Der Grund dafür ist, dass sich an diesen Übergängen die Korrelation sprunghaft auf 0 ändert, wie in Gleichung (5.32) gezeigt. Aufgrund des Terms  $(1 - 2/N_r^*)u_i^2$  im Nenner von Gleichung (5.36) hängt dieser Effekt von der konkreten Teilnehmerunsicherheit, sowie der Teilnehmerzahl ab. Ja nach Datenlage kann es daher möglich sein, dass ein als Ausreißer identifizierter Wert eine bessere Leistungsbewertung erhält als ein entsprechender Wert, der gerade nicht als Ausreißer erachtet wird. An dieser Stelle sei auch angemerkt, dass sich die Leistungsbewertung der anderen Teilnehmer ändert, wenn ein Wert als Ausreißer identifiziert wird. In einem solchen Fall erhöht sich die Korrelation der restlichen Teilnehmer mit dem Referenzwert. Die konkrete Änderung hängt auch hier von der Datenlage ab.

## 6 Methode D — Potenz-moderierter Mittelwert (Pommé & Keightley)

Mit Methode B liegt in Kapitel 4 eine unsicherheitsbasierte Auswertungsmethode mit individueller Gewichtung vor. Bei den Methoden  $C_0$  und C aus Kapitel 5 handelt es sich im Gegensatz dazu um streuungs-basierte Auswertungsmethoden mit einheitlicher Gewichtung. Die in diesem Kapitel vorgestellte Methode D vereint die genannten Eigenschaften der Methoden B und  $C_0$  und moderiert zwischen diesen, je nach Datenlage. Im Rahmen dieses Dokumentes bezeichnet Methode D die in „Determination of a reference value, associated standard uncertainty and degrees of equivalence“ von Stefaan Pommé [7] bzw. in „Determination of a reference value and its uncertainty through a power-moderated mean“ von Stefaan Pommé und John Keightley [8] veröffentlichte Auswertungsmethode von Eignungsprüfungen. Grundlage dieser Methode ist der sogenannte Mandel-Paule Mittelwert, der in „Interlaboratory evaluation of a material with unequal numbers of replicates“ [5] und „Consensus values and weighting factors“ [6] von John Mandel und Robert Paule eingeführt wurde. Im Rahmen dieses Dokumentes, wird diese Art Auswertung als Methode  $D_0$  bezeichnet.

Wie in Methode B und  $C_0/C$ , werden auch in Methode  $D_0/D$  die Referenzdaten unter der Berücksichtigung von Teilnehmerdaten berechnet, d.h. die Referenzdaten werden durch die Messungen mehrere Labore und deren anschließender statistischen Auswertung bestimmt. Ausgangspunkt dafür ist auch in diesem Kapitel der in Abschnitt 2.1 beschriebene Datensatz, d.h. es liegen die Messabweichungen  $x_i$ , sowie die dazugehörigen erweiterten Messunsicherheiten  $U_i$  der  $N$  Teilnehmer vor. Entsprechend Kapitel 2 wird weiter angenommen, dass das Ergebnis aller Teilnehmer  $i \in [1; N]$  durch eine Normalverteilung beschrieben werden kann, d.h.  $U_i = 2 \cdot u_i \quad \forall i \in [1; N]$ . Da diese Voraussetzung zentral ist für Methode B und die Berechnungsgleichungen von Methode B ein Grenzfall von Methode  $D_0/D$  sein werden, ist die Voraussetzung der Normalverteilung auch hier zentral. Des weiteren ist  $N_r$  die Anzahl der Teilnehmer in Gruppe „r“, welche die Voraussetzung erfüllen und in der Berechnung der Referenzdaten berücksichtigt werden. Da, wie in Kapitel 3 beschrieben, auch qualitative Kriterien zum Ausschluss eines Teilnehmers aus der Bestimmung der Referenzdaten führen können, ist  $N_r \leq N$ . Ein solches qualitatives Kriterium kann beispielsweise sein, dass nur nach DIN EN ISO/IEC 17025:2018 [1] akkreditierte Labore in der Berechnung berücksichtigt werden. Teilnehmer, die in der Bestimmung der Referenzdaten nicht berücksichtigt werden, sind Gruppe „n“ zugeordnet.

### 6.1 Mandel-Paule Mittelwert (Methode $D_0$ )

Wie bereits eingangs erwähnt, basiert Methode D auf dem Mandel-Paule Mittelwert, der in diesem Dokument als Methode  $D_0$  bezeichnet wird. Aus diesem Grund wird in diesem Abschnitt zunächst die Berechnung des Mandel-Paule Mittelwertes vorgestellt und beschrieben. Die Berechnungsgrundlagen aus Kapitel 4 und 5 sind an dieser Stelle vorausgesetzt.

#### 6.1.1 Berechnung der Referenzdaten

Als Grundlage des Mandel-Paule Mittelwertes können die Berechnungsgleichungen von Methode B herangezogen werden, damit ist zum einen Gleichung (4.5) für den Referenzwert

$$x_{\text{ref};B} = u_{\text{ref};B}^2 \sum_{i=1}^{N_r} \frac{x_i}{u_i^2} \quad (6.1)$$

und zum anderen Gleichung (4.3) für die Referenzunsicherheit

$$u_{\text{ref};B}^2 = \left( \sum_{i=1}^{N_r} \frac{1}{u_i^2} \right)^{-1} \quad (6.2)$$

gemeint. Der in Abschnitt 4.1.1 beschriebene  $\chi^2$ -Test, der ein wichtiger Bestandteil von Methode B ist, wird an dieser Stelle zunächst nicht weiter benötigt, taucht aber in einer modifizierten Form in Gleichung (6.6) auf. Ein Mechanismus zur Verschiebung von Teilnehmern aus Gruppe „r“ in Gruppe „n“ ist am Ende in Methode D anderweitig gegeben und in Abschnitt 6.2.2 dargestellt.

Ausgehend von den Gleichungen von Methode B, wird in Methode D<sub>0</sub> sowohl im Referenzwert, als auch in der Referenzunsicherheit ein positiver Parameter  $s$  im Nenner quadratisch zu den einzelnen Teilnehmervarianzen  $u_i^2$  addiert. Dieser Parameter  $s$  ist für alle Teilnehmer gleich. Für die Referenzdaten gilt damit

$$x_{\text{ref};D_0} = u_{\text{ref};D_0}^2 \sum_{i=1}^{N_r} \frac{x_i}{u_i^2 + s^2} \quad (6.3)$$

mit

$$u_{\text{ref};D_0}^2 = \left( \sum_{i=1}^{N_r} \frac{1}{u_i^2 + s^2} \right)^{-1} \quad (6.4)$$

Die Gewichtungsfaktoren sind demnach gegeben durch

$$w_{i;D_0} = \begin{cases} \frac{u_{\text{ref};D_0}^2}{u_i^2 + s^2} & \text{Teilnehmer in Gruppe „r“,} \\ 0 & \text{Teilnehmer in Gruppe „n“.} \end{cases} \quad (6.5)$$

Aus der Kombination der Gleichungen (6.4) und (6.5) wird klar, dass die Gewichtungsfaktoren die in Gleichung (2.31) geforderte Normierung erfüllen. Zur Bestimmung des Parameters  $s$  wird die Hilfsgröße

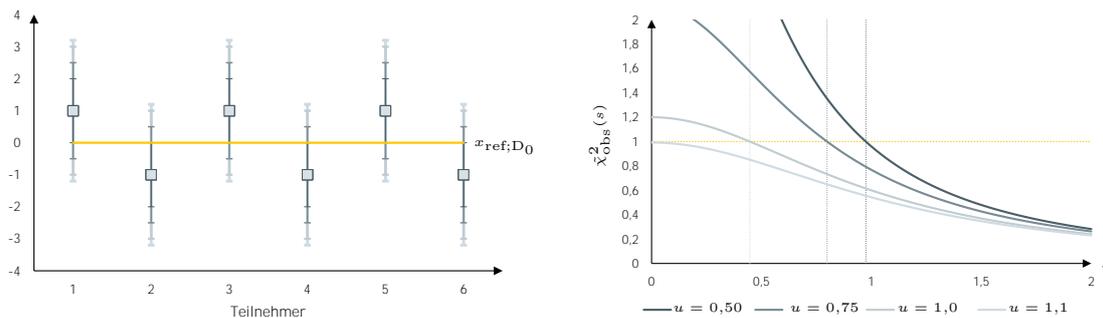
$$\tilde{\chi}_{\text{obs}}^2(s) = \frac{1}{N_r - 1} \sum_{i=1}^{N_r} \frac{(x_i - x_{\text{ref};D_0})^2}{u_i^2 + s^2} \quad (6.6)$$

definiert. Der Parameter  $s \geq 0$  ist kleinstmöglich so zu wählen, dass

$$\tilde{\chi}_{\text{obs}}^2(s) \leq 1, \quad (6.7)$$

wodurch  $s$  eindeutig bestimmt ist und damit auch Gleichungen (6.3) und (6.4). Gleichung (6.7) kann sukzessive numerisch gelöst werden, indem als Startwert  $s = 0$  verwendet wird. Falls damit  $\tilde{\chi}_{\text{obs}}^2 > 1$  ist, so ist der Parameter solange schrittweise zu erhöhen, bis die Hilfsgröße den Wert eins annimmt. Es sei darauf hingewiesen, dass durch  $x_{\text{ref};D_0}$  in Gleichung (6.6) auch die Gleichungen (6.3) und (6.4) in die Bestimmung von  $s$  einfließen. Die Hilfsgröße  $\tilde{\chi}_{\text{obs}}^2$  erinnert in ihrer Definition an eine Mischung

zwischen der Variable  $\chi_{\text{obs}}^2$  des  $\chi^2$ -Test in Methode B, vgl. Gleichung (4.7), und der empirischen Standardabweichung aus Methode C<sub>0</sub>, vgl. Gleichung (5.4). Dadurch wird bereits angedeutet, dass durch den Parameter  $s$  das Zusammenspiel von Unsicherheit und Streuung berücksichtigt wird. Die konkrete Wahl der Hilfsgröße  $\tilde{\chi}_{\text{obs}}^2$  wird später noch einmal deutlich, wenn die Grenzfälle des Parameters  $s$  und der Methode D<sub>0</sub> im Allgemeinen analysiert werden.



**Abbildung 6.1**

Die im Bild links gezeigten alternierenden Teilnehmerwerte  $x_i = \pm x$  mit  $x = 1$  ergeben aufgrund ihrer Symmetrie immer den Referenzwert  $x_{\text{ref};D_0} = 0$ , unabhängig davon, wie die Unsicherheiten  $u_i = u \forall i \in [1; N_r]$  ausfallen und wie der  $s$ -Parameter gewählt wird. Rechts ist der Verlauf der Hilfsgröße  $\tilde{\chi}_{\text{obs}}^2$  in Abhängigkeit von  $s$  dargestellt. Je kleiner die Unsicherheit  $u$ , desto größer ist  $s$ , jedoch nicht größer als  $N_r / (N_r - 1)$  im gewählten Beispiel. Der Vergleich beider Abbildungen zeigt, dass der Parameter  $s$  die Tatsache berücksichtigt, dass durch die Unsicherheiten  $u_i = u \forall i \in [1; N_r]$  die Streuung der Werte  $x_i$  nicht abgedeckt bzw. erklärbar ist.

Um den Parameter  $s$  und dessen Bestimmung zu veranschaulichen, wird nachfolgend ein alternierender Beispieldatensatz gewählt, d.h.  $x_{i=2n-1} = x$  und  $x_{i=2n} = -x$  mit  $n \in [1; N_r/2]$  für  $N_r = 6$ , der in Abbildung 6.1 veranschaulicht ist. Aufgrund der Symmetrie der Teilnehmerdaten  $x_i$  ist der Referenzwert  $x_{\text{ref};D_0} = 0$ , unabhängig davon, welchen Wert der Parameter  $s$  annimmt. Entsprechend ist die quadratische Differenz  $(x_i - x_{\text{ref};D_0})^2$  in der Hilfsgröße aus Gleichung (6.6) konstant und betragsmäßig gleich  $x$ . Der Einfachheit halber sei zudem die Unsicherheit  $u_i$  aller Teilnehmer gleich, d.h.  $u_i = u \forall i \in [1; N_r]$ . Damit lässt sich Gleichung (6.7) analytisch lösen und der Parameter  $s$  ist festgelegt durch

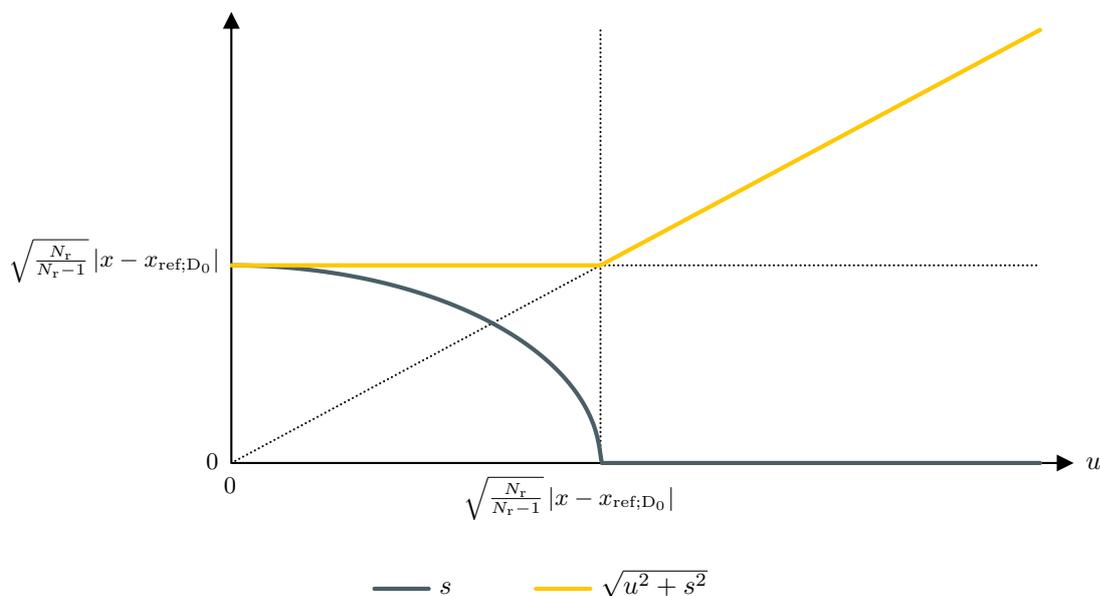
$$s = \begin{cases} \sqrt{\frac{N_r}{N_r-1} (x - x_{\text{ref};D_0})^2 - u^2} & u \leq \sqrt{\frac{N_r}{N_r-1}} |x - x_{\text{ref};D_0}|, \\ 0 & \text{sonst.} \end{cases} \quad (6.8)$$

Damit wird ausgedrückt, was in Abbildung 6.1 dargestellt ist. Die Unsicherheiten  $u_i = u$  müssen hinreichend groß sein und die vorliegende Streuung der Werte  $x_i = \pm x$  hinreichend abdecken. Ist das nicht der Fall, so wird der Parameter  $s$  größer als Null sein.

Im genannten Beispiel wird für Gleichung (6.8) deutlich, dass die Quadratsumme zwischen Unsicherheit  $u$  und Parameter  $s$  eine Konstante liefert, falls  $s > 0$  ist

$$\sqrt{u^2 + s^2} = \begin{cases} \sqrt{\frac{N_r}{N_r-1}} |x - x_{\text{ref};D_0}| & s > 0, \\ u & s = 0. \end{cases} \quad (6.9)$$

Für  $s > 0$  resultiert für das Datenbeispiel demnach die empirische Stichprobenstandardabweichung. Das zeigt sich auch in Abbildung 6.2. Je größer die Unsicherheit  $u$  wird, desto kleiner wird der Parameter  $s$ , die Quadratsumme aus beidem bleibt jedoch konstant, solange  $s > 0$  ist. Ist die Unsicherheit hinreichend groß, d.h.  $u \geq \sqrt{\frac{N_r}{N_r-1}} |x - x_{\text{ref};D_0}|$ , so wird der Parameter  $s$  nicht weiter benötigt, d.h.  $s = 0$ , da die Streuung der Werte durch die Unsicherheiten hinreichend abgedeckt ist. Entsprechend der Hilfsgröße  $\tilde{\chi}_{\text{obs}}^2$  wird im genannten Beispiel der Wert  $\sqrt{\frac{N_r}{N_r-1}} |x - x_{\text{ref};D_0}|$ , d.h. die empirische Stichprobenstandardabweichung der vorliegenden Daten, als untere Grenze für die Unsicherheit erachtet. Wird die Grenze durch die Teilnehmerunsicherheiten nicht erreicht, so wird der Parameter  $s$  ergänzt. An dieser Stelle sei darauf hingewiesen, dass diese einfachen Zusammenhänge der gewählten Datenlage im genannten Beispiel geschuldet sind. Im Allgemeinen kann die Wahl des Parameters  $s$  nicht analytisch ermittelt werden. Das Verhalten im gezeigten Beispiel unterstreicht jedoch zum einen den Charakter des Parameters  $s$  und zum anderen wird an dieser Stelle bereits das Zusammenspiel zwischen Teilnehmerwertestreuung und Teilnehmerunsicherheiten deutlich.



**Abbildung 6.2**

Gezeigt ist der Verlauf des Parameters  $s$  in Abhängigkeit von der Unsicherheit  $u$  im Beispiel von Abbildung 6.1. Je größer die Unsicherheit  $u$  wird, desto kleiner wird der Parameter  $s$ , allerdings bleibt die Wurzel der Quadratsumme aus  $u$  und  $s$  solange konstant, bis  $s = 0$  ist und ist dann nur noch durch  $u$  gegeben. Die kritischen Punkte sind durch  $\sqrt{\frac{N_r}{N_r-1}} |x - x_{\text{ref};D_0}|$  gegeben, d.h. die Hilfsgröße  $\tilde{\chi}_{\text{obs}}^2$  erachtet gemäß den vorliegenden Daten  $\sqrt{\frac{N_r}{N_r-1}} |x - x_{\text{ref};D_0}|$  als untere Grenzen für die Unsicherheit. Ist dieser Mindestwert durch die Teilnehmerunsicherheiten nicht gegeben, wird der Parameter  $s$  ergänzt.

An dieser Stelle können nun die beiden Grenzfälle des Mandel-Paule Mittelwertes aus Gleichung (6.3) und

dessen Unsicherheit (6.4) betrachtet werden. Der erste Grenzfall für  $s = 0$  ist offensichtlich, da dieser bereits zu Beginn als Motivation verwendet wurde. Für  $s = 0$ , d.h. wenn die Teilnehmerunsicherheiten im Hinblick auf die vorliegende Streuung der Werte  $x_i$  hinreichend groß sind, entspricht Methode  $D_0$  dem gewichteten Mittelwert aus Methode B. Konkret bedeutet das für die Referenzunsicherheit

$$\begin{aligned}
 u_{\text{ref};D_0}^2 &= \left( \sum_{i=1}^{N_r} \frac{1}{u_i^2 + s^2} \right)^{-1} \\
 &\xrightarrow{s=0} \left( \sum_{i=1}^{N_r} \frac{1}{u_i^2} \right)^{-1} \\
 &= u_{\text{ref};B}^2.
 \end{aligned} \tag{6.10}$$

Für den Referenzwert folgt entsprechend

$$\begin{aligned}
 x_{\text{ref};D_0} &= u_{\text{ref};D_0}^2 \sum_{i=1}^{N_r} \frac{x_i}{u_i^2 + s^2} \\
 &\xrightarrow{s=0} u_{\text{ref};B}^2 \sum_{i=1}^{N_r} \frac{x_i}{u_i^2} \\
 &= x_{\text{ref};B}.
 \end{aligned} \tag{6.11}$$

Im Grenzfall  $s = 0$  verhält sich Methode  $D_0$  wie Methode B, d.h. es liegt eine unsicherheitsbasierte Auswertung mit individueller Gewichtung über die Teilnehmervarianzen vor.

Der zweite Grenzfall ist für  $u_i = 0 \forall i \in [1; N_r]$  gegeben. Zwar tritt dieser Fall in der Realität nicht ein, das Verhalten dieses Grenzfall es zeigt sich jedoch auch dann, wenn  $s \gg u_i \forall i \in [1; N_r]$ , d.h. die berichteten Unsicherheiten der Teilnehmer sind deutlich kleiner als die vorliegende Streuung der Teilnehmerwerte. In diesem Fall gilt für die Referenzunsicherheit

$$\begin{aligned}
 u_{\text{ref};D_0}^2 &= \left( \sum_{i=1}^{N_r} \frac{1}{u_i^2 + s^2} \right)^{-1} \\
 &\xrightarrow{u_i=0} \left( \sum_{i=1}^{N_r} \frac{1}{s^2} \right)^{-1} \\
 &= \frac{s^2}{N_r} = \frac{1}{N_r(N_r - 1)} \sum_{i=1}^{N_r} (x_i - x_{\text{ref};D_0})^2 \\
 &= u_{\text{ref};C_0;1}^2
 \end{aligned} \tag{6.12}$$

Für den Referenzwert gilt

$$\begin{aligned}
 x_{\text{ref};D_0} &= u_{\text{ref};D_0}^2 \sum_{i=1}^{N_r} \frac{x_i}{u_i^2 + s^2} \\
 &\xrightarrow{u_i=0} u_{\text{ref};B}^2 \sum_{i=1}^{N_r} \frac{x_i}{s^2} \\
 &= \frac{1}{N_r} \sum_{i=1}^{N_r} x_i \\
 &= x_{\text{ref};C_0}.
 \end{aligned} \tag{6.13}$$

In diesem Grenzfall resultieren der arithmetische Mittelwert und dessen empirische Standardabweichung. Das bedeutet, dass falls die Teilnehmerunsicherheiten im Vergleich zur vorliegenden Streuung der Teilnehmerwerte verhältnismäßig klein sind, verhält sich Methode  $D_0$  wie die streuungsbasierte Auswertung nach Methode  $C_0$ , in der eine einheitliche Gewichtung vorliegt.

Anhand der beiden Grenzfälle wird klar, dass der Parameter  $s$  je nach Datenlage zwischen einer unsicherheitsbasierten und einer streuungsbasierten Auswertung, sowie zwischen der individuellen Teilnehnergewichtung über die Unsicherheiten und der einheitlichen Gewichtung des arithmetischen Mittelwertes moderiert. Je nach Datenlage wird das Verhalten der genannten Grenzfälle beobachtet, aber auch Kompromisse und Zwischenfälle können abgebildet werden. Insbesondere ist es auch möglich, dass der Referenzwert ähnlich zu Methode B ist, die Referenzunsicherheit aber eher Methode C entspricht und umgekehrt. Anhand von Beispieldaten wird in Abschnitt 6.1.3 die Interpolation und Moderation zwischen diesen Grenzfällen durch den Parameter  $s$  verdeutlicht.

### 6.1.2 Leistungsbewertung

In diesem Abschnitt wird die Leistungsbewertung in Methode  $D_0$  näher betrachtet. Zunächst wird hierfür festgehalten, dass es sich, entsprechend den Gleichungen (6.3), (6.4) und (6.5), bei Methode  $D_0$  um ein lineares Modell gemäß Abschnitt 2.2.4 handelt. Der  $E_{n;i}$ -Wert ist demnach durch Gleichung (2.37) gegeben, d.h.

$$E_{n;i} = \frac{x_i - x_{\text{ref};D_0}}{2\sqrt{(1 - 2w_{i;D_0})u_i^2 + u_{\text{ref};D_0}^2}}. \tag{6.14}$$

Einsetzen der Gewichtungsfaktoren aus Gleichung (6.5) liefert

$$E_{n;i} = \begin{cases} \frac{x_i - x_{\text{ref};D_0}}{2\sqrt{\left(1 - 2\frac{u_{\text{ref};D_0}^2}{u_i^2 + s^2}\right)u_i^2 + u_{\text{ref};D_0}^2}} & \text{Teilnehmer } i \text{ in Gruppe „r“,} \\ \frac{x_i - x_{\text{ref};D_0}}{2\sqrt{u_i^2 + u_{\text{ref};D_0}^2}} & \text{Teilnehmer } i \text{ in Gruppe „n“.} \end{cases} \tag{6.15}$$

Im Grenzfall  $s = 0$  resultiert der  $E_{n;i}$ -Wert von Methode B aus Gleichung (4.11). Falls  $s \gg u_i \forall i \in [1; N_r]$ , so ergibt sich der  $E_{n;i}$ -Wert von Methode  $C_0$  in Gleichung (5.8). Die gleichen Grenzfälle zeigen sich auch

in der Kovarianz, bzw. dem Korrelationskoeffizienten. Für die Kovarianz gilt gemäß Gleichung [2.35]

$$u(x_i; x_{\text{ref};D_0}) = \begin{cases} \frac{u_{\text{ref};D_0}^2 u_i^2}{u_i^2 + s^2} & \text{Teilnehmer } i \text{ in Gruppe „r“,} \\ 0 & \text{Teilnehmer } i \text{ in Gruppe „n“.} \end{cases} \quad (6.16)$$

Der Korrelationskoeffizient ist entsprechend Gleichung [2.38] für Teilnehmer  $i$  in Gruppe „r“ gegeben durch

$$r_{i;\text{ref};D_0} = \frac{u_{\text{ref};D_0} u_i}{u_i^2 + s^2} \rightarrow \begin{cases} \frac{u_{\text{ref};D_0}}{u_i} & s = 0, \\ \frac{u_i}{N_r u_{\text{ref};D_0}} & s \gg u_i \text{ mit } s = N_r u_{\text{ref};D_0}^2 \text{ gemäß Gleichung [6.12],} \end{cases} \quad (6.17)$$

sodass sich auch hier die Grenzfälle von Methode B ( $s = 0$ ) und Methode  $C_0$  ( $s \gg u_i$ ) ergeben. Da es sich bei Methode  $D_0$  eindeutig um ein lineares Modell im Sinne von Abschnitt 2.2.4 handelt, ist bereits durch die dortigen Ausführungen klar, dass der Korrelationskoeffizient wohl definiert ist mit  $0 \leq r_{i;\text{ref};D_0} < 1$ .

### 6.1.3 Charakteristik der Methode

In Abschnitt 6.1.1 wurden bereits wesentliche Eigenschaften des Parameters  $s$ , sowie dessen Zusammenspiel mit der Streuung der Teilnehmerwerte  $x_i$  und den Teilnehmerunsicherheiten  $u_i$  anhand eines einfachen Beispiels dargestellt. In diesem Abschnitt werden nun weitere Eigenschaften der Methode  $D_0$  aufgezeigt und insbesondere die Moderation durch den Parameter  $s$  zwischen den Grenzfällen von Methode B und  $C_0$  verdeutlicht. Hierfür wird der in Abbildung 6.3 gezeigte Beispieldatensatz definiert

$$\begin{aligned} \mathbf{x} &= \{10,0; 12,0; 8,0; 9,0; 6,0\}, \\ u(\mathbf{x}) &= \{0,50; 1,0; 1,5; 2,0; 2,5\}. \end{aligned} \quad (6.18)$$

Der Datensatz wurde so gewählt, dass sowohl der  $\chi^2$ -Test in Methode B bestanden ist, als auch Methode C keinen Ausreißer identifiziert. Neben den Teilnehmerdaten sind in Abbildung 6.3 auch die resultierenden Referenzdaten der Methoden B,  $C_0$ , C und  $D_0$  gezeigt. Um die Eigenschaften von Methode  $D_0$  herauszuarbeiten, ist es sinnvoll den Vergleich zu den Referenzdaten der Methoden B und  $C_0/C$  zu ziehen. Mit den Ausführungen in Kapitel 4 ist klar, dass Methode B maßgeblich durch Teilnehmer mit verhältnismäßig kleinen Unsicherheiten dominiert ist. Im gezeigten Beispiel zeigt sich dies daran, dass der Referenzwert im Wesentlichen durch den Wert von Teilnehmer 1 gegeben ist. Die Referenzunsicherheit ist ebenfalls erwartungsgemäß kleiner als die kleinste Teilnehmerunsicherheit, die in diesem Beispiel für Teilnehmer 1 vorliegt. Konkret sind die Referenzdaten in Methode B gegeben durch

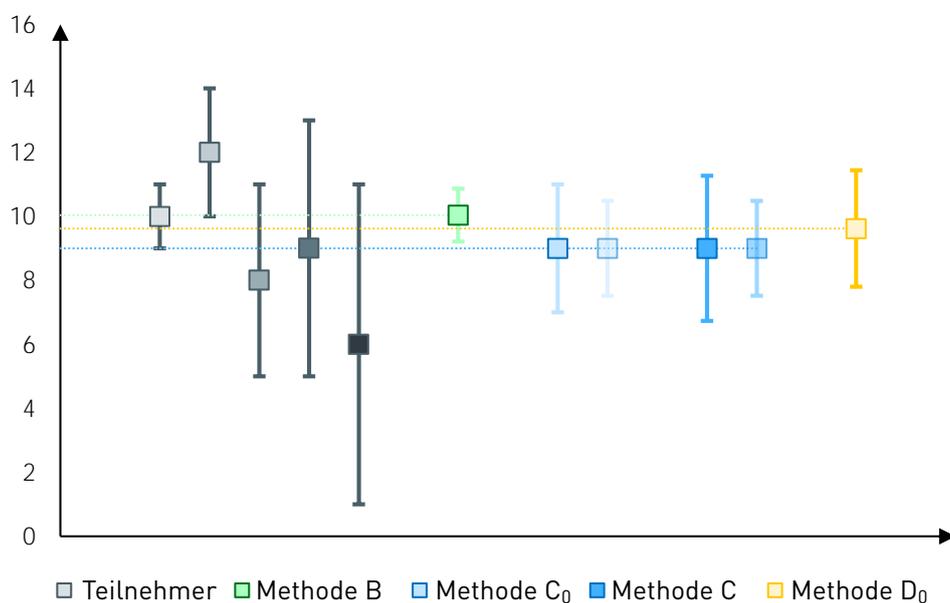
$$x_{\text{ref};B} \simeq 10,0 \quad \text{und} \quad u_{\text{ref};B} \simeq 0,41. \quad (6.19)$$

Da die Beispieldaten so gewählt wurden, dass der Algorithmus von Methode C keinen Wert als Ausreißer identifiziert, unterscheiden sich die Referenzdaten von Methode  $C_0$  und C lediglich in der Referenzunsicherheit, aufgrund des Faktors 1,134 aus Gleichung [5.19]. Da die Streuung der Werte  $x_i$  größer ist als

die Fehlerfortpflanzung der Teilnehmerunsicherheiten ( $u_{\text{ref};C_0/C;2} \simeq 0,74$ ), ist die Unsicherheit letztlich durch die empirische Standardabweichung aus Gleichung (5.4) bzw. (5.21) gegeben. Die Referenzdaten lauten:

$$\begin{aligned} x_{\text{ref};C_0} &= 9,0 & \text{und} & & u_{\text{ref};C_0} &= 1,0 \\ x_{\text{ref};C} &= 9,0 & \text{und} & & u_{\text{ref};C} &= 1,134 \end{aligned} \quad (6.20)$$

Vor dem Hintergrund dieser Grenzfälle kann nun Methode  $D_0$  näher analysiert werden. Bevor die resultierenden Referenzdaten, die bereits in Abbildung 6.3 gezeigt sind, näher betrachtet werden, ist ein Blick auf die Gewichtungsfaktoren  $w_{i;\text{ref};D_0}$  aus Gleichung (6.5) hilfreich und in Abbildung 6.4 gezeigt.

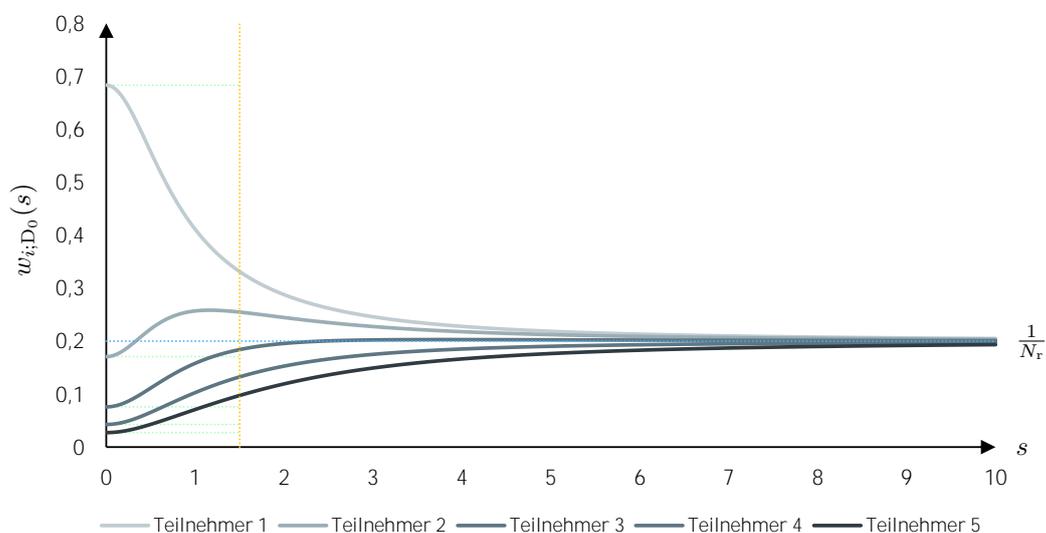


**Abbildung 6.3**

Gezeigt sind die in Gleichung (6.18) definierten Beispieldaten der Teilnehmer, sowie die daraus resultierenden Referenzdaten nach Methode B,  $C_0$ , C und  $D_0$ . Die jeweils zweiten Werte für Methode  $C_0$  und C zeigen die Unsicherheit der Fehlerfortpflanzung. Da diese kleiner ist als die Streuung der Teilnehmerwerte ist die Referenzunsicherheit durch die empirische Standardabweichung gegeben. Methode B ist im Wesentlichen durch Teilnehmer 1 dominiert, aufgrund der kleinen Unsicherheit. Methode  $D_0$  erscheint als Kompromiss zwischen Methode B und den Methoden  $C_0/C$ , wobei der Referenzwert näher an Methode B, die Referenzunsicherheit jedoch näher an Methode  $C_0$  liegt.

Für  $s = 0$  resultiert die individuelle Gewichtung von Methode B. In dieser Gewichtung über die Teilnehmervarianzen wird Teilnehmer 1 deutlich stärker gewichtet als alle anderen Teilnehmer. Durch die Ergänzung des Parameters  $s$  in Gleichung (6.5), wird die Gewichtung der einzelnen Teilnehmer vereinheitlicht, indem letztlich durch den Parameter  $s$  die Unsicherheit der einzelnen Teilnehmer nach oben korrigiert wird. Im Verlauf der Gewichtungsfaktoren aus Abbildung 6.4 zeigt sich, dass durch den Parameter  $s$  insbesondere die Gewichtung von Teilnehmer 1 stark verringert wird, wohingegen die Gewichtung der Teilnehmer 1 bis 4 zunimmt. Da der Parameter  $s$  für alle Teilnehmer gleichermaßen ergänzt wird, schneiden sich die Verläufe der Gewichtungsfaktoren nicht, d.h. qualitativ wird ein Teilneh-

mer  $i$ , der in Methode B stärker gewichtet ist als ein anderer Teilnehmer  $j$ , auch in Methode  $D_0$  stärker gewichtet. Lediglich das konkrete Verhältnis der Gewichtungsfaktoren zueinander wird angepasst, d.h. vereinheitlicht. Der konkrete Verlauf ist stets vom vorliegenden Datensatz abhängig, wobei hier lediglich die Teilnehmerunsicherheiten relevant sind. Die Summe aller Gewichtungsfaktoren ist zu jedem Zeitpunkt gleich 1, da die Normierung aus Gleichung (2.31) stets erfüllt ist. Mit größer werdendem Parameter  $s$  zeigt sich eine Konvergenz aller Gewichtungsfaktoren gegen die einheitliche Gewichtung  $1/N_r$  des arithmetischen Mittelwertes aus Methode  $C_0$ . Methode  $D_0$  moderiert in Abhängigkeit von  $s$  somit zwischen der individuellen Gewichtung von Methode B und der einheitlichen Gewichtung der Methoden  $C_0$ .

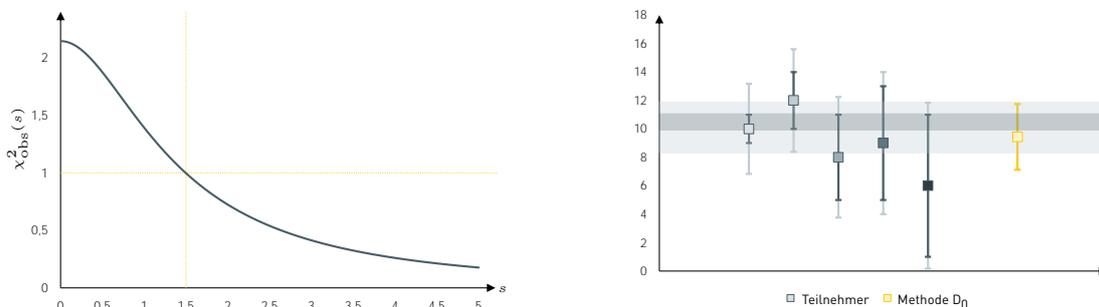


#### Abbildung 6.4

Gezeigt ist der Verlauf der in Gleichung (6.5) definierten Gewichtungsfaktoren von Methode  $D_0$  in Abhängigkeit vom Parameter  $s$ . Für  $s = 0$  liegt die Gewichtung von Methode B vor und jeder Teilnehmer wird individuell über die vorliegende Teilnehmervarianz gewichtet. Je größer der Parameter  $s$  wird, desto einheitlicher wird die Gewichtung aller Teilnehmer und eine Konvergenz hin zur einheitlichen Gewichtung nach Methode  $C_0$  ist zu sehen ( $N_r = 5$  im gewählten Beispiel). Da der Parameter  $s$  letztlich gemäß Gleichung (6.7) nicht willkürliche Werte annimmt, ist die tatsächliche Gewichtung der Teilnehmer durch den Wert bei  $s = 1,5$  im vorliegenden Beispiel gegeben. An diesem Punkt ist zu sehen, dass insbesondere der Einfluss von Teilnehmer 1 in Methode  $D_0$  im Vergleich zu Methode B deutlich reduziert ist.

Da der konkrete Wert des Parameters  $s$  letztlich durch Gleichung (6.7) festgelegt ist, resultiert für Methode  $D_0$  eine Art vereinheitlichte individuelle Gewichtung. Wie in Abbildung 6.5 zu sehen ist, nimmt im vorliegenden Beispieldatensatz aus Gleichung (6.18) der Parameter  $s$  den Wert  $s = 1,5$  an. Für diesen Wert zeigt sich in Abbildung 6.4, dass zwar nach wie vor jeder Teilnehmer anders, d.h. individuell gewichtet wird und insbesondere Teilnehmer 1 und 2 stärker gewichtet werden als die Teilnehmer 3, 4 und 5. Allerdings sind die Unterschiede in der Gewichtung nicht mehr so groß wie in Methode B, d.h. die individuelle Gewichtung ist zu einem gewissen Grad vereinheitlicht. Der Grad der Vereinheitlichung hängt vom finalen Parameter  $s$  und somit von der Datenlage ab. Die in Abbildung 6.4 beispielhaft darge-

stellte Vereinheitlichung der Gewichtungsfaktoren von Methode  $D_0$  bedeutet im Modell von Methode B effektiv eine Vergrößerung der Teilnehmerunsicherheiten  $u_i$ , wie Abbildung 6.5 zeigt. Zwar wird zu allen Teilnehmerunsicherheiten  $u_i$  der gleiche Wert  $s$  quadratisch addiert, die Änderung der effektiven Unsicherheit ist jedoch für Teilnehmer, die eine verhältnismäßig kleine Unsicherheit berichten größer. In Abbildung 6.5 ist beispielhaft der Überdeckungsbereich der erweiterten Teilnehmerunsicherheiten mit und ohne  $s$ -Parameter markiert. Da die Streuung der Teilnehmerdaten nicht vollständig durch die berichteten Teilnehmerunsicherheiten  $u_i$  erklärt werden kann, liegt für  $s = 0$  eine Inkonsistenz der Daten vor, die anhand des schmalen dunkelgrauen Streifens in Abbildung 6.5 veranschaulicht ist. Der Parameter  $s$  wird letztlich so lange vergrößert, bis über die effektiven Unsicherheiten  $u_i^2 + s^2$  eine hinreichende Konsistenz vorliegt und die Streuung der Daten hinreichend abgedeckt ist (vgl. hellgrauer Streifen in Abbildung 6.5).



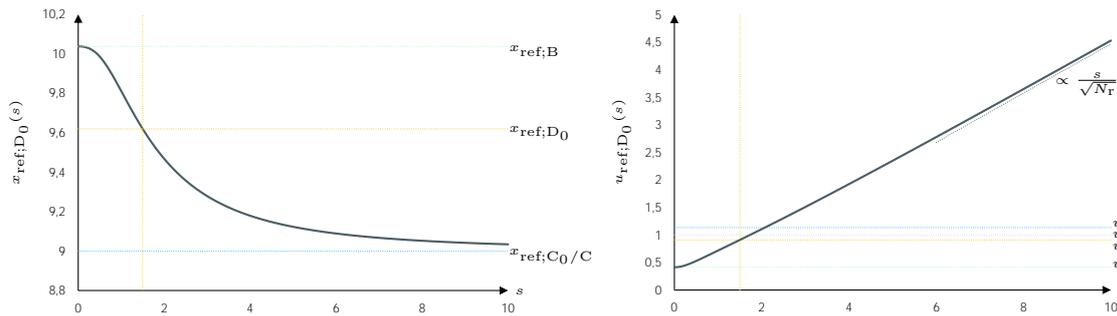
**Abbildung 6.5**

*Links: Für den Beispieldatensatz aus Gleichung (6.18) ist der Verlauf der Hilfsgröße  $\chi_{\text{obs}}^2$ , die zur Bestimmung des Parameters  $s$  dient gezeigt. Im konkreten Beispiel ist  $s = 1,5$ .*

*Rechts: Im Modell von Methode B bewirkt der Parameter  $s$  in Methode  $D_0$  effektiv die Vergrößerung der einzelnen Teilnehmerunsicherheiten, wodurch die Inkonsistenz der Teilnehmerdaten behoben wird.*

Wie in Gleichung (6.3) und (6.4) gezeigt, sind in Methode  $D_0$  letztlich der Referenzwert und die Referenzunsicherheit durch den Parameter  $s$  maßgeblich beeinflusst. Der Verlauf ist in Abbildung 6.6 dargestellt. Vor allem im Referenzwert zeigt sich die Moderation zwischen Methode B und Methode  $C_0/C$  durch den Parameter  $s$ , die in ähnlicher Weise bereits für die Gewichtungsfaktoren in Abbildung 6.4 zu erkennen ist, da im linearen Modell nach Gleichung (2.29) die Gewichtungsfaktoren direkt in die Berechnung des Referenzwertes einfließen. Der konkrete Verlauf zwischen den Grenzfällen von Methode B und Methode  $C_0$  hängt vom vorliegenden Datensatz ab und kann hier nur exemplarisch für die Daten aus Gleichung (6.18) gezeigt werden. Da, wie in Abbildung 6.5 gezeigt, im vorliegenden Datensatz der Parameter  $s$  durch 1,5 gegeben ist, resultiert in Methode  $D_0$  ein Referenzwert, der zwischen den Referenzwerten von Methode B und  $C_0/C$ , jedoch etwas näher am Referenzwert aus Methode B liegt.

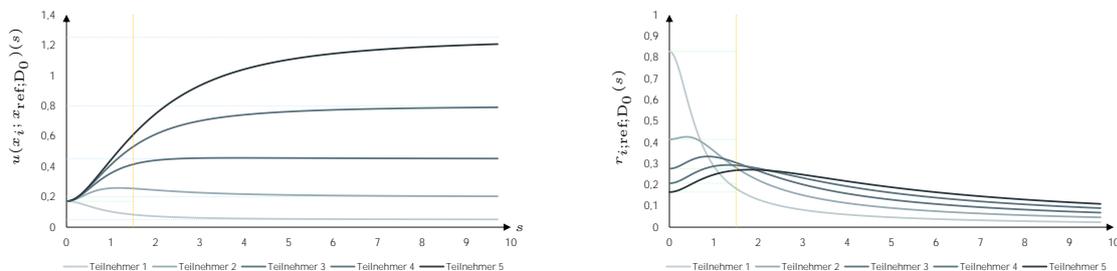
Der Verlauf der Referenzunsicherheit in Abhängigkeit vom Parameter  $s$  zeigt, wie Abbildung 6.6 veranschaulicht, ein etwas anderes Verhalten als der Referenzwert und die Gewichtungsfaktoren. Zunächst wird auch für  $s = 0$  die Referenzunsicherheit von Methode B reproduziert. Für größer werdendes  $s$  zeigt sich jedoch, dass die Referenzunsicherheit nach Gleichung (6.4) prinzipiell beliebig groß werden kann, aufgrund der in Gleichung (6.12) gezeigten Proportionalität  $u_{\text{ref};D_0} \propto s/\sqrt{N_r}$ . Grund dafür ist, dass in der Referenzunsicherheit von Methode  $D_0$  noch die Normierung der Faktoren  $1/(u_i^2 + s^2)$  fehlt, die dann



**Abbildung 6.6**

Links ist der Verlauf der Referenzunsicherheit  $x_{\text{ref};D_0}$  in Methode  $D_0$  in Abhängigkeit vom Parameter  $s$  zu sehen und rechts der Verlauf der dazugehörigen Referenzunsicherheit  $u_{\text{ref};D_0}$ . Am Startwert  $s = 0$  liegen die Referenzdaten aus Methode B vor. Der Referenzwert konvergiert für größer werdendes  $s$  gegen den Referenzwert aus Methode  $C_0/C$ . Der konkrete Verlauf zwischen den beiden Grenzfällen hängt von der Datenlage ab. Die Referenzunsicherheit kann in Abhängigkeit vom Parameter  $s$  prinzipiell beliebig groß werden. Da  $s$  jedoch nicht beliebig ist, ist der tatsächliche Wert für  $s = 1,5$  erreicht und liegt für das gezeigte Beispiel zwischen den Unsicherheiten von Methode B und  $C_0/C$ .

in den Gewichtungsfaktoren und somit auch im Referenzwert vorliegt. In der Anwendung ist  $s$  jedoch letztlich durch Gleichung (6.7) festgelegt und wird nicht beliebig groß. Im vorliegenden Datenbeispiel resultiert eine Referenzunsicherheit, die zwischen den Werten von Methode B und  $C_0$  liegt. Im Vergleich zwischen Referenzwert und Referenzunsicherheit zeigt sich in Abbildung 6.6, dass der Referenzwert von Methode  $D_0$  näher am Referenzwert von Methode B ist, die Referenzunsicherheit von Methode  $D_0$  ist jedoch näher an der Referenzunsicherheit von Methode  $C_0$ . Demnach liegt selbst in der Kombination aus Referenzwert und Referenzunsicherheit eine Art Kompromiss zwischen Methode B und  $C_0$  vor.

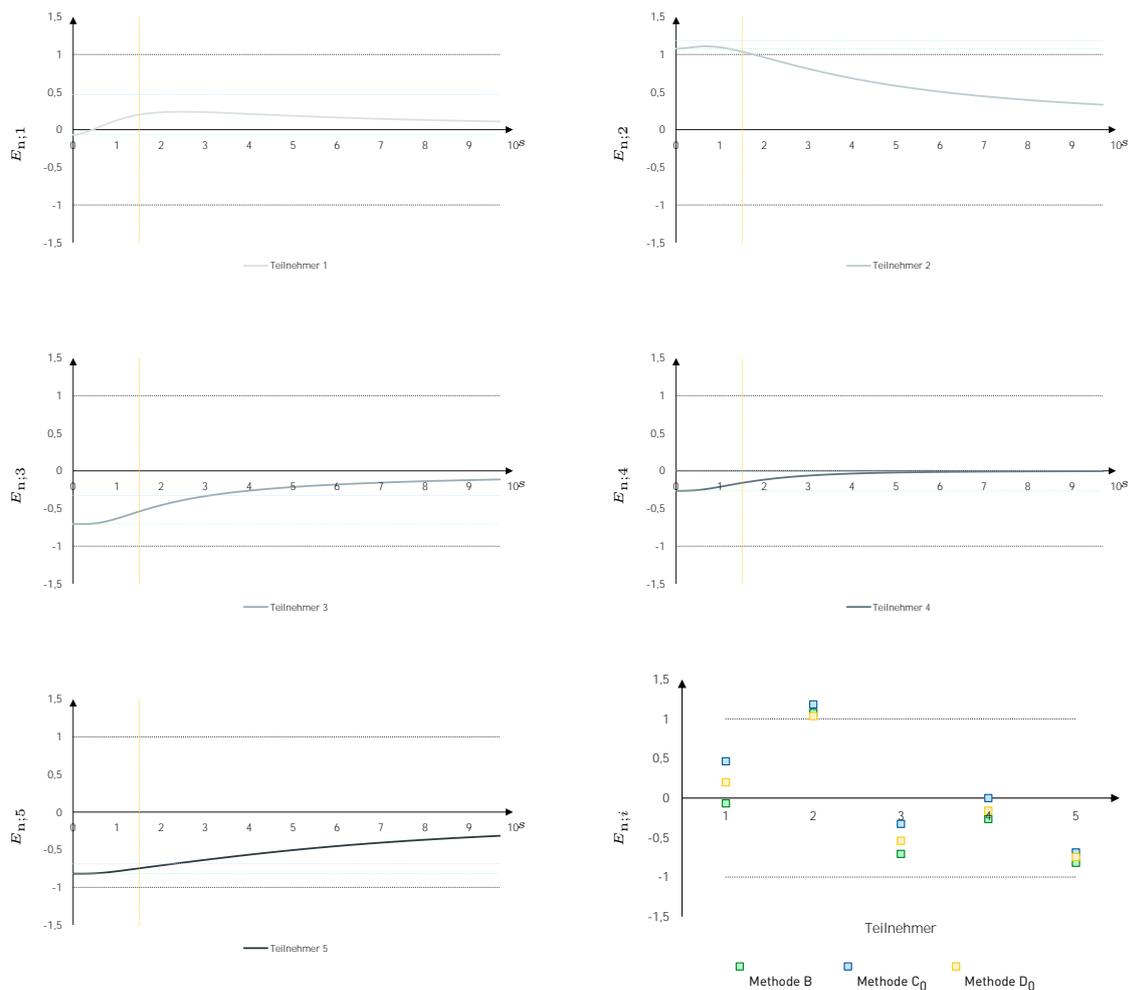


**Abbildung 6.7**

Links ist der Verlauf der Kovarianz und rechts der Verlauf des Korrelationskoeffizienten in Abhängigkeit vom Parameter  $s$  gezeigt. Für die Kovarianz zeigen sich die diskutierten Grenzfälle von Methode B und Methode  $C_0$ . Für den Korrelationskoeffizienten zeigt sich aufgrund des Verhaltens der Referenzunsicherheit ein Konvergenz gegen Null für sehr große Parameter  $s$ . Für  $s = 0$  resultieren die Werte von Methode B.

Ergänzend zu den bisherigen Diskussionen ist in Abbildung 6.7 der Verlauf der Kovarianz und des Korrelationskoeffizienten gezeigt. Für die Kovarianz zeigen sich die bereits diskutierten Grenzfälle von Methode B und Methode  $C_0$ , da die Kovarianz im linearen Modell nach Gleichung (2.35) direkt proportional zu den Gewichtungsfaktoren ist. Der Korrelationskoeffizient hingegen reproduziert nur den Grenzfall von Methode B für  $s = 0$ . Für größer werdendes  $s$  konvergieren die Korrelationskoeffizienten aller Teilnehmer gegen Null. Der Grund hierfür ist der Verlauf der Referenzunsicherheit und die Tatsache,

dass  $u_{\text{ref};D_0}/(u_i^2 + s^2)$  im Gegensatz zu  $u_{\text{ref};D_0}^2/(u_i^2 + s^2)$  keine Normierungsbedingung erfüllt. Interessant ist im vorliegenden Datenbeispiel, dass beim finalen Parameter  $s = 1,5$  der Korrelationskoeffizient von Teilnehmer 1 kleiner ist als der Korrelationskoeffizient der anderen Teilnehmer, obgleich für  $s = 0$  der Wert von Teilnehmer 1 eine sehr starke Korrelation mit dem Referenzwert von Methode B aufweist. Insbesondere zeigt sich im Vergleich zum Verhalten der Gewichtungsfaktoren aus Abbildung 6.4, dass sich die Verläufe der Korrelationskoeffizienten der Teilnehmer schneiden können.



**Abbildung 6.8**

Gezeigt ist die Leistungsbewertung der einzelnen Teilnehmer in Methode  $D_0$  in Abhängigkeit vom Parameter  $s$ . Startpunkt ist für  $s = 0$  stets der  $E_{n;i}$ -Wert von Methode B. Der weitere Verlauf hängt vom Teilnehmer individuell ab. Ein  $E_{n;i}$ -Wert zwischen dem von Methode B und Methode  $C_0$  resultiert in 4 von 5 Fällen. Für größer werdende Werte von  $s$  konvergieren alle Leistungsbewertungen in Methode  $D_0$  gegen Null, da die Referenzunsicherheit, wie in Abbildung 6.6 beliebig groß werden kann. Dieses Verhalten wird in der Anwendung jedoch durch Gleichung (6.7) verhindert

Abschließend kann noch ein Blick auf die Leistungsbewertung der einzelnen Teilnehmer geworfen werden. In Abbildung 6.8 sind der Verlauf des  $E_{n;i}$ -Wertes für die einzelnen Teilnehmer in Methode  $D_0$  in Abhängigkeit vom Parameter  $s$ , sowie die Grenzfälle der Leistungsbewertung gemäß Methode

B und  $C_0$  gezeigt. In allen Fällen resultiert für  $s = 0$  der  $E_{n;i}$ -Wert nach Methode B. Der Grenzfall von Methode  $C_0$  wird in der Leistungsbewertung nach Methode  $D_0$  nicht immer erreicht bzw. für einen  $s$ -Wert angenommen (vgl. Teilnehmer 1 und 2). Für größer werdendes  $s$  konvergieren alle  $E_{n;i}$ -Werte gegen Null, was am Verlauf der Referenzunsicherheit liegt die letztlich beliebig groß werden kann. Abbildung 6.8 zeigt, dass zwischen der Leistungsbewertung von Methode B und Methode  $C_0$  durchaus Unterschiede vorliegen können. Allerdings ist die finale Leistungsbewertung nach Methode  $D_0$  nicht immer in der Mitte zwischen der Leistungsbewertung von Methode B und  $C_0$  zu finden. Hier zeigt sich, dass das finale Zusammenspiel von Referenzwert, Referenzunsicherheit und Korrelation zu einer individuellen Leistungsbewertung nach Methode  $D_0$  führt, die in vielen Fällen eine Art Kompromiss zwischen den Auswertungsmethoden B und  $C_0$  ist, in manchen Fällen aber auch davon abweichende Ergebnisse liefert.

Anmerkung: Das gezeigte Verhalten ist auf den Beispieldatensatz aus Gleichung (6.18) beschränkt, obgleich daran viele allgemeine Eigenschaften von Methode  $D_0$  sichtbar werden. Die konkrete Abhängigkeit vom Parameter  $s$ , sowie die durch diesen Parameter eingeführte Moderation zwischen den Grenzfällen von Methode B und  $C_0$  hängt prinzipiell vom vorliegenden Datensatz ab. So können beispielsweise je nach Datenlage für die Referenzunsicherheit von Methode  $D_0$  auch Werte resultieren, die etwas größer sind als die Referenzunsicherheit von Methode B und Methode  $C_0$ . Auch im Referenzwert kann der Verlauf für manche  $s$ -Werte so sein, dass  $x_{\text{ref};D_0}(s)$  nicht mehr zwischen den Grenzwerten von Methode B und Methode  $C_0$  liegt. Weitere Datenbeispiele sind in Kapitel 7 zu finden.

## 6.2 Potenz-moderierter Mittelwert (Methode D)

Die in [7, 8] veröffentlichte Methode des potenz-modierten Mittelwertes wird im Rahmen dieses Dokumentes als Methode D bezeichnet und ist Gegenstand der nachfolgenden Betrachtungen. Die in den vorherigen Abschnitten dargestellte Methode  $D_0$ , die den Mandel-Paule Mittelwert bezeichnet, und ihre Eigenschaften sind der Ausgangspunkt und die Grundlage von Methode D. Insbesondere werden die vereinheitlichte individuelle Gewichtung, sowie der streuungs- und unsicherheitsbasierte Auswertungsansatz von Methode  $D_0$  in Methode D aufgegriffen und verstärkt.

### 6.2.1 Berechnung der Referenzdaten

Wie in Abschnitt 6.1 gezeigt, basiert der Mandel-Paule Mittelwert aus Methode  $D_0$  auf dem gewichteten Mittelwert von Methode B, wobei ein zusätzlicher Parameter  $s$  eingeführt wird, der letztlich zwischen Methode B und Methode  $C_0$ , je nach Datenlage, moderiert. Davon ausgehend, werden in Methode D zwei weitere Parameter ergänzt,  $\alpha$  und  $S$ . Diese fließen in den Gewichtungsfaktor der Methode D gemäß

$$w_{i;D} = \begin{cases} \frac{u_{\text{ref};D}^2}{S^{2-\alpha} \sqrt{u_i^2 + s^2}^\alpha} & \text{Teilnehmer in Gruppe „r“}, \\ 0 & \text{Teilnehmer in Gruppe „n“} \end{cases} \quad (6.21)$$

ein. Für die Referenzdaten bedeutet dies

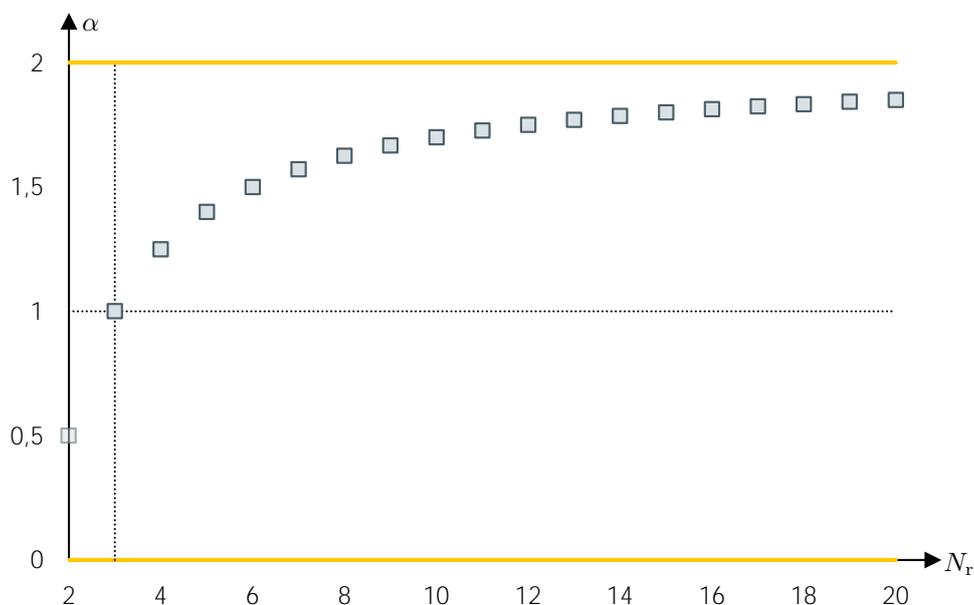
$$x_{\text{ref};D} = \frac{u_{\text{ref};D}^2}{S^{2-\alpha}} \sum_{i=1}^{N_r} \frac{x_i}{\sqrt{u_i^2 + s^{2\alpha}}} \quad (6.22)$$

mit

$$u_{\text{ref};D}^2 = S^{2-\alpha} \left( \sum_{i=1}^{N_r} \frac{1}{\sqrt{u_i^2 + s^{2\alpha}}} \right)^{-1}. \quad (6.23)$$

Durch die Gleichungen (6.21), (6.22) und (6.23) sind die Voraussetzungen des linearen Modells aus Abschnitt 2.2.4 erfüllt und die Summe der Gewichtungsfaktoren ist auf 1 normiert. Da der Parameter  $\alpha$  im Exponenten steht, wird der Referenzwert aus Gleichung (6.22) als potenz-moderierter Mittelwert bezeichnet. Der Wertebereich des Parameters  $\alpha$  ist zunächst gegeben durch

$$\alpha \in [0; 2]. \quad (6.24)$$



#### Abbildung 6.9

Gezeigt ist der Wertebereich des Parameters  $\alpha$  gemäß den Gleichungen (6.24) und (6.25). Der Parameter  $\alpha$  kann entweder händisch auf 0 oder 2 gesetzt werden, oder entsprechend Gleichung (6.25) bestimmt werden. Ist letzteres der Fall, so ist der minimale Wert durch 1 gegeben, da sinnvollerweise  $N_r \geq 3$ , wie in Kapitel 8 gezeigt. Für größer werdendes  $N_r$  zeigt sich eine Konvergenz gegen 2.

Die Wahl des Parameters  $\alpha$  hängt davon ab, ob die berichteten Unsicherheiten  $u_i$  im Vergleich zu den berichteten Abweichungen  $x_i$  als hinreichend konsistent erachtet werden. Standardmäßig kann

$$\alpha = 2 - \frac{3}{N_r} \quad (6.25)$$

verwendet werden. Damit ist der Parameter  $\alpha$  insbesondere abhängig von der Teilnehmerzahl  $N_r$ , d.h.

der Anzahl an Daten, die in die Bestimmung der Referenzdaten einfließen. Je kleiner  $N_r$  ist, d.h. je geringer die Datenlage, desto kleiner ist  $\alpha$ . Da  $N_r \geq 3$  sein sollte, wie in Kapitel 8 gezeigt wird, ist der kleinste Wert, den der Parameter  $\alpha$  nach Gleichung (6.25) annehmen kann, durch 1 gegeben. Für eine größer werdende Teilnehmerzahl  $N_r$ , konvergiert der Parameter  $\alpha$  gegen 2, wie in Abbildung 6.9 dargestellt. Neben Gleichung (6.25) kann der Parameter jedoch auch auf 0 oder 2 gesetzt werden. Dabei entspricht  $\alpha = 0$  dem Fall, dass in den Daten eine Diskrepanz vorliegt und die beobachteten Abweichungen  $x_i$  nicht durch die berichteten Unsicherheiten  $u_i$  abgedeckt und erklärbar sind. Sind die berichteten Unsicherheiten  $u_i$  hingegen aussagekräftig und decken die berichteten Abweichungen  $x_i$  ab, d.h. es liegen konsistente Daten vor, so kann  $\alpha = 2$  gewählt werden. In diesem Fall entspricht Methode D dem Mandel-Paule Mittelwert aus Methode D<sub>0</sub>. Gleichung (6.25) entspricht hingegen einer mäßig inkonsistenten Datenlage in der die Unsicherheiten tendenziell unterschätzt sind, aber dennoch mit den berichteten Abweichungen übereinstimmen. Die nachfolgende Tabelle zeigt eine Zusammenfassung der Wahl des Parameters  $\alpha$ . Da sowohl die Lage der Abweichungen  $x_i$ , als auch die Größenordnung der berichteten Unsicherheiten  $u_i$  berücksichtigt wird, handelt es sich bei Methode D um eine streuungs- und unsicherheitsbasierte Auswertungsmethode.

$$\alpha = 0$$

- Die Unsicherheiten  $u_i$  sind nicht informativ und können die Streuung der Abweichungen  $x_i$  nicht erklären.
- Es liegen inkonsistente Daten vor.
- Grenzfall: Methode C<sub>0</sub> (arithmetischer Mittelwert)

$$\alpha = 2 - \frac{3}{N_r} \in [1; 2)$$

- Informative Unsicherheiten  $u_i$ , die jedoch tendenziell unterschätzt sind.
- Für geringe Teilnehmerzahlen  $N_r$ , wird die unzureichende Datengrundlage berücksichtigt.

$$\alpha = 2$$

- Die Unsicherheiten  $u_i$  sind realistisch und in Übereinstimmung mit der Streuung der berichteten Messwerte  $x_i$ .
- Es liegen konsistente Daten vor.
- Grenzfall: Methode D<sub>0</sub>/B (Mandel-Paule Mittelwert/gewichteter Mittelwert)

In Abschnitt 6.1 wurde bereits gezeigt, dass der Parameter  $s$  zwischen dem arithmetischen Mittelwert von Methode C<sub>0</sub> und dem gewichteten Mittelwert von Methode B moderiert. In Methode D wird diese Moderation durch den Parameter  $\alpha$  erweitert. Wie der obigen Tabelle zu entnehmen ist, liegt für  $\alpha = 0$  der Grenzfall des arithmetischen Mittelwertes und für  $\alpha = 2$  der Grenzfall des Mandel-Paule Mittelwertes vor. Ersterer wird später genauer betrachtet, da hierfür noch der Parameter  $S$  von Bedeutung ist. Für  $\alpha = 2$  hingegen ist aus den Gleichungen (6.22) und (6.23) klar, dass  $S^{2-\alpha} = 1$  ist und der Parameter  $S$  somit entfällt. In diesem Fall werden die Referenzdaten aus Methode D<sub>0</sub> reproduziert. Mehr noch, da

$\alpha = 2$  nur dann gesetzt wird, wenn die Daten konsistent sind, kann erwartet werden, dass der Parameter  $s$  des Mandel-Paule Mittelwertes gleich 0 ist und somit letztlich der gewichtete Mittelwert aus Methode B reproduziert wird. Umgekehrt kann ein Parameter  $s > 0$  im Mandel-Paule Mittelwert ein Hinweis darauf sein, dass der Parameter  $\alpha < 2$  ist.

Insgesamt ist in Methode D das Zusammenspiel der beiden Parameter  $S$  und  $\alpha$  charakteristisch, weshalb die Auswirkung der beiden Parameter nur zusammen betrachtet werden kann. Nachdem beschrieben wurde, wie der Parameter  $\alpha$  zu wählen ist, ist nachfolgend die Bestimmungsgleichung des Parameters  $S$  gezeigt, es gilt

$$S = \sqrt{N_r \cdot \max(u_{\text{ref};C_0;1}^2; u_{\text{ref};D_0}^2)}. \quad (6.26)$$

Hier ist zunächst festzustellen, dass es sich bei  $u_{\text{ref};C_0;1}^2$  nach Gleichung (5.4) und  $u_{\text{ref};D_0}^2$  nach Gleichung (6.4) um Varianzen eines Mittelwertes handelt. Durch die Definition des Parameters  $S$  in Gleichung (6.26) wird klar, dass es sich bei  $S$  um eine Stichprobenstandardabweichung handelt, d.h. um eine Art vereinheitlichte Teilnehmerunsicherheit. Wie in den obigen Ausführungen des Parameters  $\alpha$  bereits aufgezeigt, moderiert dieser zwischen Methode  $C_0$  ( $\alpha = 0$ ) und Methode  $D_0$  ( $\alpha = 2$ ). Diese beiden Grenzfälle werden auch im Parameter  $S$  aufgegriffen, zum einen durch die empirische Standardabweichung des arithmetischen Mittelwertes  $u_{\text{ref};C_0;1}$  und zum anderen durch die Standardunsicherheit des Mandel-Paule Mittelwertes  $u_{\text{ref};D_0}$ . Wie ein Vergleich der Gleichungen (6.22) und (6.23) zeigt, kann der Parameter  $S$  im Referenzwert gekürzt werden, d.h.  $S$  hat keinen Einfluss auf die Lage des Referenzwertes, sondern lediglich auf die Größe der Referenzunsicherheit. Durch die Definition des Parameters  $S$  in Gleichung (6.26) wird erneut unterstrichen, dass mit Methode D, wie bereits mit Methode  $D_0$  eine streuungs- und unsicherheitsbasierte Auswertungsmethode vorliegt.

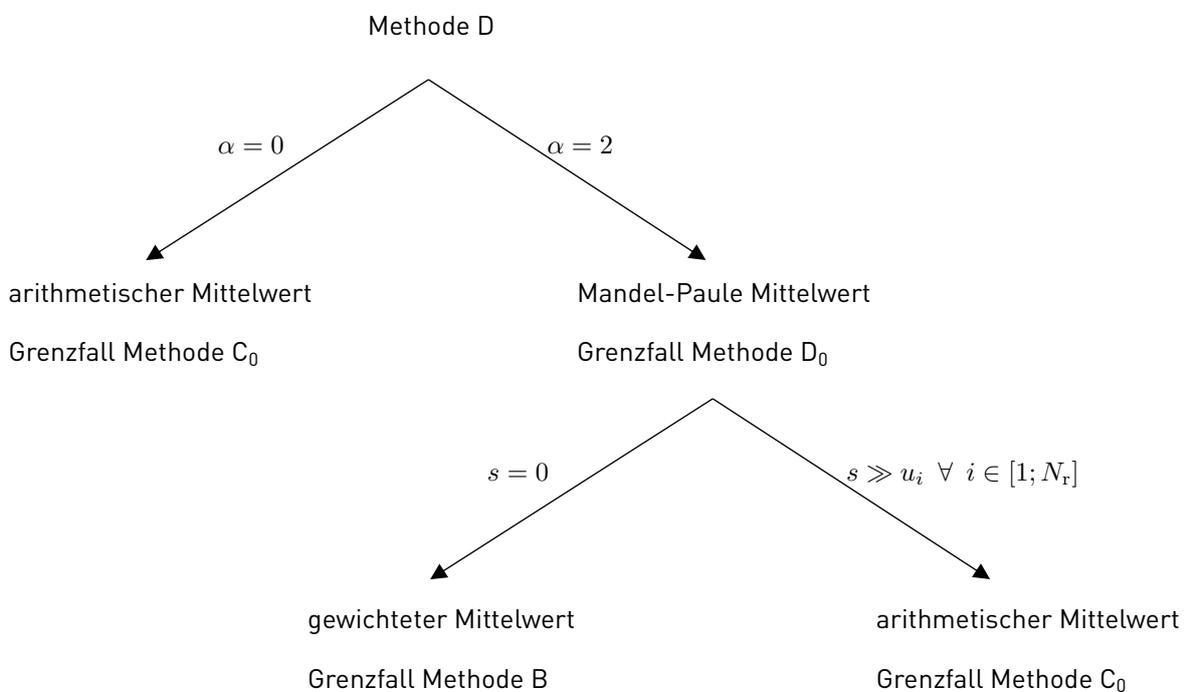
Um das Zusammenspiel der Parameter  $\alpha$  und  $S$  zu verstehen, ist es hilfreich den Summand eines Teilnehmers  $i$  aus Gruppe „r“ in der Referenzunsicherheit nach Gleichung (6.23) näher zu betrachten:

$$\underbrace{\frac{1}{S^{2-\alpha}}}_{\text{einheitliches Gewicht}} \cdot \underbrace{\frac{1}{\sqrt{u_i^2 + s^{2\alpha}}}}_{\text{(vereinheitlichtes) individuelles Gewicht}} \quad (6.27)$$

Wie Gleichung (6.27) zeigt, setzt sich der Summand in der Referenzunsicherheit in Methode D aus einem einheitlichen und einem (vereinheitlichten) individuellen Anteil zusammen. Letzterer stammt aus dem Mandel-Paule Mittelwert, bei dem die individuelle Gewichtung aus Methode B durch den Parameter  $s$  ggf. bereits vereinheitlicht wurde. Wie in Abschnitt 6.1 jedoch gezeigt ist, wird durch den Parameter  $s$ , je nach Datenlage, nur eine teilweise Vereinheitlichung der individuellen Gewichtung erzielt. Eine weitere Vereinheitlichung wird in Methode D durch den Parameter  $\alpha$  vorgenommen. Dieser taucht in beiden Faktoren aus Gleichung (6.27) auf und kann als eine Art Gewichtsstück verstanden werden, das zwischen der Waagschale der einheitlichen und der Waagschale der (vereinheitlichten) individuellen Gewichtung ausgleicht. Das was an individueller Gewichtung weggenommen wird, wird durch den Parameter  $S$  an einheitlicher Gewichtung aufgefüllt. Für  $\alpha = 2$  wird alles in die Waagschale der (vereinheitlichten) individuellen Gewichtung gelegt, da  $S^{2-\alpha} = 1$  ist. Für  $\alpha = 0$  wird alles in die Waagschale

der einheitlichen Gewichtung und die individuelle Gewichtung verschwindet, da  $\sqrt{u_i^2 + s^{2\alpha}} = 1$  ist.

Wichtig ist zu verstehen, dass sich in Methode D das Zusammenspiel der drei Parameter  $s$ ,  $S$  und  $\alpha$  niederschlägt, wobei vor allem  $s$  und  $\alpha$  zwischen der einheitlichen Gewichtung von Methode  $C_0$  und der individuellen Gewichtung von Methode B moderieren. Durch den Parameter  $s$  werden, wie in Abschnitt 6.1 gezeigt, die Teilnehmerunsicherheiten vergrößert. Da dies für alle Teilnehmer durch den gleichen Parameter  $s$  geschieht, werden die Unsicherheiten dadurch auch, je nach Datenlage mehr oder weniger, vereinheitlicht. Mit  $\alpha$  liegt in Methode D ein zusätzlicher Moderator zwischen individueller und einheitlicher Gewichtung vor, der, wie in Gleichung (6.25) gezeigt, auch die Teilnehmeranzahl, d.h. die Größe der Datengrundlage berücksichtigt. Schematisch ergeben sich die in Abbildung 6.10 dargestellten Grenzfälle. Insbesondere können aber nicht nur die Grenzfälle, sondern beliebige Zwischenbereiche realisiert werden.



#### Abbildung 6.10

Dargestellt sind die Grenzfälle, die sich in Methode D durch die Parameter  $\alpha$  und  $s$  ergeben. Durch die Moderation der beiden Parameter können insbesondere auch die verschiedenen Zwischenbereiche abgebildet werden.

### 6.2.2 Leistungsbewertung und Ausschluss von Teilnehmerdaten

Die Leistungsbewertung nach Methode D kann direkt aus den Überlegungen des linearen Modells in Abschnitt 2.2.4 übernommen werden. Der  $E_{n,i}$ -Wert ist demnach durch Gleichung (2.37) gegeben, d.h.

$$E_{n;i} = \frac{x_i - x_{\text{ref};D}}{2\sqrt{(1 - 2w_{i;D})u_i^2 + u_{\text{ref};D}^2}}. \quad (6.28)$$

Mit den Gewichtungsfaktoren aus Gleichung (6.21) folgt

$$E_{n;i} = \begin{cases} \frac{x_i - x_{\text{ref};D}}{2\sqrt{\left(1 - 2\frac{u_{\text{ref};D}^2}{S^{2-\alpha}\sqrt{u_i^2 + s^{2\alpha}}}\right)u_i^2 + u_{\text{ref};D}^2}} & \text{Teilnehmer } i \text{ in Gruppe „r“,} \\ \frac{x_i - x_{\text{ref};D}}{2\sqrt{u_i^2 + u_{\text{ref};D}^2}} & \text{Teilnehmer } i \text{ in Gruppe „n“.} \end{cases} \quad (6.29)$$

Die im vorherigen Abschnitt angesprochenen Grenzfälle ergeben sich auch im  $E_{n;i}$ -Wert. Für die Kovarianz gilt gemäß Gleichung (2.35)

$$u(x_i; x_{\text{ref};D}) = \begin{cases} \frac{u_{\text{ref};D}u_i^2}{S^{2-\alpha}\sqrt{u_i^2 + s^{2\alpha}}} & \text{Teilnehmer } i \text{ in Gruppe „r“,} \\ 0 & \text{Teilnehmer } i \text{ in Gruppe „n“.} \end{cases} \quad (6.30)$$

Der Korrelationskoeffizient ist entsprechend Gleichung (2.38) für Teilnehmer  $i$  in Gruppe „r“ gegeben durch

$$r_{i;\text{ref};D} = \frac{u_{\text{ref};D}u_i}{S^{2-\alpha}\sqrt{u_i^2 + s^{2\alpha}}} \quad (6.31)$$

Der Korrelationskoeffizient ist gemäß den Ausführungen für das lineare Modell in Abschnitt 2.2.4 wohl definiert mit  $0 \leq r_{i;\text{ref};D} < 1$ .

Über die Definition des  $E_{n;i}$ -Wertes ist entsprechend [7, 8] ein Ausschlusskriterium für extreme Teilnehmerwerte definiert. Zunächst werden aus allen in Gruppe „r“ vorliegenden Teilnehmerdaten die Referenzdaten gemäß den Gleichungen (6.22) und (6.23) berechnet. Anschließend werden für alle Teilnehmer in Gruppe „r“ die  $E_{n;i}$ -Werte gemäß Gleichung (6.29) berechnet. Darauf basierend wird folgendes Ausschlusskriterium definiert

$$w_{i;D} \equiv 0 \quad \Leftrightarrow \quad E_{n;i} > c \quad \text{mit } c \in [1; 2]. \quad (6.32)$$

Alle Teilnehmer aus Gruppe „r“, deren Leistungen anhand des  $E_{n;i}$ -Wertes negativ bewertet werden und den kritischen Grenzwert  $c \in [1; 2]$  überschreiten, sind von Gruppe „r“ in Gruppe „n“ zu verschieben, d.h. aus der Bestimmung der Referenzdaten auszuschließen. Der kritische Wert  $c$  in Gleichung (6.32) kann dabei je nach Anforderungen festgelegt werden. Die untere Grenze ist durch 1 gegeben, für die obere Grenze empfiehlt sich der Wert 2. Standardmäßig wird

$$c = 1,25 \quad (6.33)$$

empfohlen. Mit den verbleibenden Teilnehmern in Gruppe „r“ ist dann die Bestimmung der Referenzdaten neu durchzuführen und das Ausschlusskriterium erneut auszuwerten. Der Ablauf wird solange wiederholt, bis alle Teilnehmer in Gruppe „r“, die in die Bestimmung der Referenzdaten einfließen, das Ausschlusskriterium aus Gleichung (6.32) bestehen.

Anmerkung 1: Das hier formulierte Ausschlusskriterium basiert auf [7, 8], weicht aber von der dort gezeigten Vereinfachung ab. Der Grund dafür ist, dass ein Fehler in der Vereinfachung vermutet wird, da diese so nur für Methode B anwendbar ist, nicht aber für Methode D. Die in [7, 8] angegebenen Gleichungen sind weniger ausreißersensitiv als jenes aus Gleichung (6.32).

Anmerkung 2: Das an dieser Stelle für Methode D eingeführte Ausschlusskriterium kann auch für Methode  $D_0$  angewandt werden.

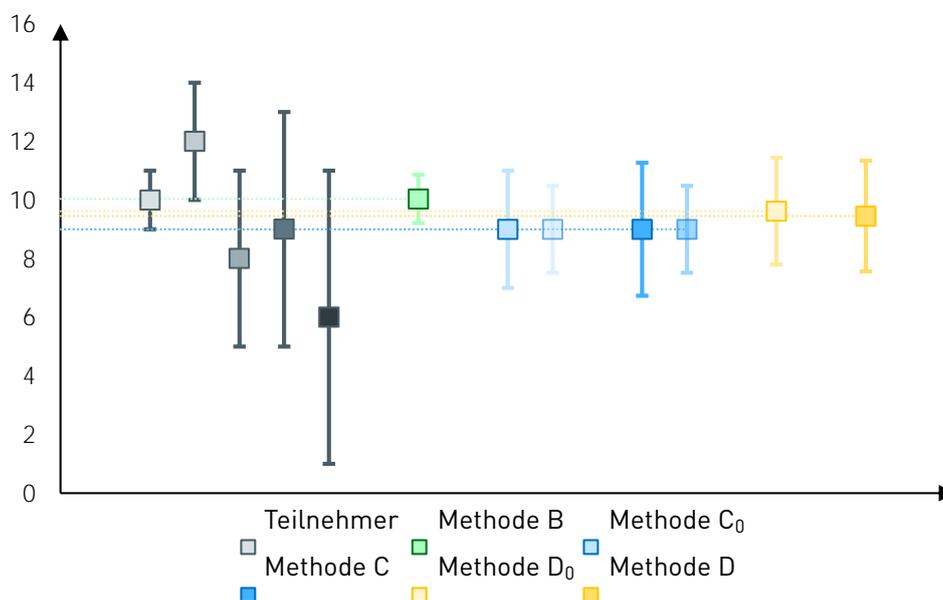
### 6.2.3 Charakteristik der Methode

Um die Eigenschaften von Methode D und die Auswirkungen der Parameter  $\alpha$  und  $S$  nachfolgend darzustellen, wird der Beispieldatensatz aus Gleichung (6.18) aufgegriffen. Dieser ist, zusammen mit den Referenzdaten aller Methoden in Abbildung 6.11 dargestellt. Methode D liefert konkret

$$x_{\text{ref};D} \simeq 9,5 \quad \text{und} \quad u_{\text{ref};D} \simeq 0,95. \quad (6.34)$$

Zum Vergleich, Methode  $D_0$  ergibt

$$x_{\text{ref};D_0} \simeq 9,6 \quad \text{und} \quad u_{\text{ref};D_0} \simeq 0,91. \quad (6.35)$$

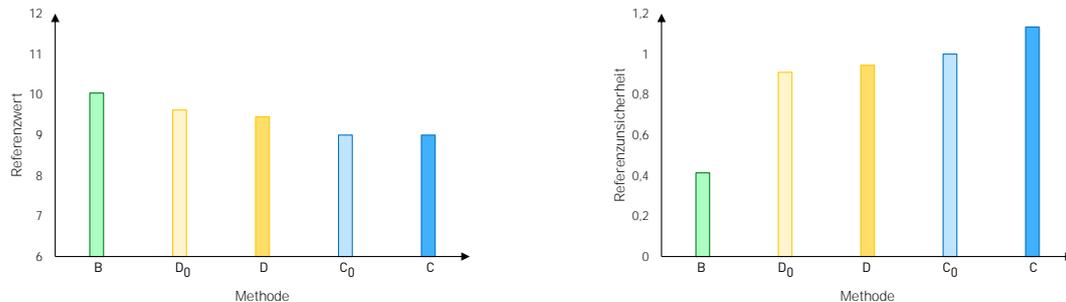


**Abbildung 6.11**

Gezeigt ist der in Gleichung (6.18) definierte Beispieldatensatz, der bereits in Abbildung 6.3 dargestellt ist. Ergänzend sind hier die Referenzdaten von Methode D gezeigt.

Abbildung 6.12 zeigt den Vergleich der Referenzdaten in den verschiedenen Auswertungsmethoden. Obleich der Unterschied zwischen Methode  $D_0$  und D für den betrachteten Beispieldatensatz minimal ist, zeigt sich, dass Methode D näher an Methode  $C_0/C$  ist, als Methode  $D_0$ . Dieses Verhalten ist eine

Konsequenz des zusätzlichen Parameters  $\alpha$ , der nach Gleichung (6.25) im betrachteten Beispiel den Wert  $\alpha = 1,4$  annimmt. Da  $\alpha < 2$ , wird die Gewichtung in Methode D weiter vereinheitlicht, wodurch Methode D näher an Methode  $C_0$  heranrückt. Der Parameter  $S$  ist im vorliegenden Beispiel durch  $S \simeq 2,24$  gegeben, und entspricht der empirischen Stichprobenstandardabweichung, da  $u_{\text{ref};C_0;1} > u_{\text{ref};D_0}$ .



**Abbildung 6.12**

Für den Beispieldatensatz aus Gleichung (6.18) sind die Referenzdaten in den verschiedenen Auswertungsmethoden gezeigt. Im Vergleich zwischen Methode  $D_0$  zeigt sich zwar, dass die Ergebnisse sehr ähnlich sind, Methode D jedoch näher an Methode  $C_0/C$  liegt. Dies liegt an der zusätzlichen Vereinheitlichung der Teilnehmergewichtungen durch die Parameter  $\alpha$  und  $S$  (im Beispiel:  $\alpha = 1,4$  und  $S \simeq 2,24$ ).

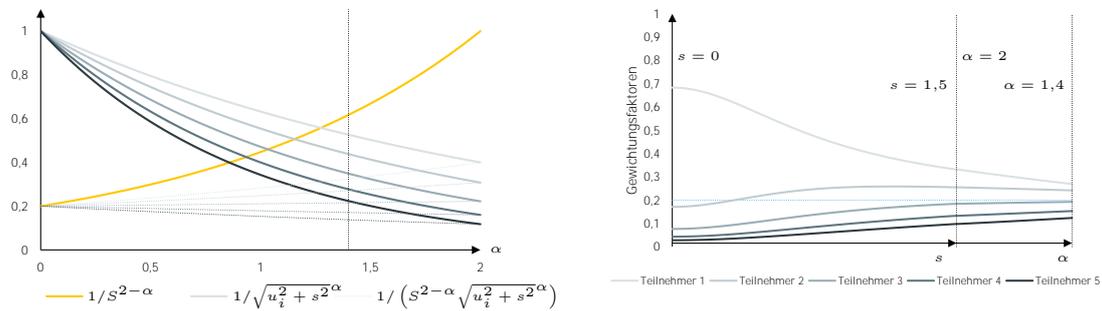
Wie in Abschnitt 6.2.1 beschrieben, stellt der Parameter  $S$  eine vereinheitlichte Teilnehmerunsicherheit dar. Der Parameter  $\alpha$  moderiert zwischen den beiden Faktoren aus Gleichung (6.27). Der Verlauf der beiden Komponenten, d.h. des einheitlichen Gewichtungsanteils  $1/S^{2-\alpha}$  und des (vereinheitlicht) individuellen Gewichtungsanteils  $1/\sqrt{u_i^2 + s^{2\alpha}}$  ist in Abbildung 6.13 in Abhängigkeit vom Parameter  $\alpha$  gezeigt. Es zeigt sich, dass der einheitliche und der individuelle Faktor der Gewichtung ein gegenläufiges Verhalten in Abhängigkeit von  $\alpha$  aufweisen. Darin spiegelt sich die Moderation des Parameters  $\alpha$  zwischen der einheitlichen und der individuellen Gewichtungskomponente wider. Wie Abbildung 6.13 zeigt, setzt der Parameter  $\alpha$  dadurch die durch den Parameter  $s$  begonnene Vereinheitlichung fort.

Ergänzend zu Abbildung 6.13 zeigt Abbildung 6.14 die finalen Gewichtungsfaktoren der verschiedenen Methoden im Vergleich. Für alle Teilnehmer zeigt sich, dass die Gewichtung nach Methode  $D_0$  und D sich ähnlich sind, was bereits durch die Referenzdaten in Abbildung 6.12 impliziert ist. Allerdings zeigt sich erneut, dass die Parameter  $\alpha$  und  $S$  die Gewichtungsfaktoren aus Methode  $D_0$  in Richtung derer aus Methode  $C_0$  abändern, d.h. vereinheitlichen. Dies kann bedeuten, dass die Gewichtungsfaktoren aus Methode  $D_0$  vergrößert oder verkleinert werden.

Im Modell der Gewichtungsfaktoren von Methode B aus Gleichung (4.4), können für die anderen Methoden effektive Teilnehmerunsicherheiten definiert werden

$$u_{i;\text{eff};m} = \frac{u_{\text{ref};m}}{\sqrt{w_{i;m}}} \quad \text{mit } m \in \{B; C_0; C; D_0; D\}. \quad (6.36)$$

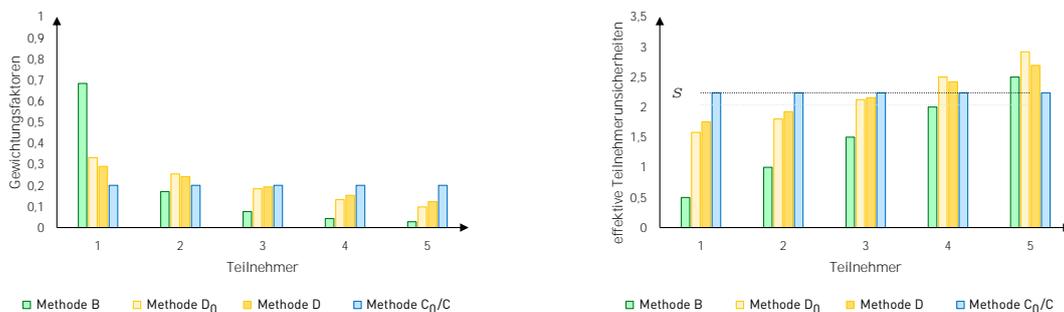
Abbildung 6.14 zeigt diese effektiven Teilnehmerunsicherheiten für die jeweiligen Methoden dargestellt. Auch hier zeigt sich, dass die Methoden  $D_0$  und D zwischen Methode B und Methode  $C_0$  moderieren.



**Abbildung 6.13**

Links ist der Verlauf der einheitlichen ( $1/S^{2-\alpha}$ ) und der individuellen ( $1/\sqrt{u_i^2 + s^{2\alpha}}$ ) Gewichtungskomponente aus Gleichung (6.27), sowie deren Produkt gezeigt. Das gegenläufige Verhalten der Komponenten spiegelt die Moderation durch den Parameter  $\alpha$  wider. Die Abbildung rechts zeigt den Verlauf der vollständigen Gewichtungsfaktoren in Methode D. Ausgehend von der Gewichtung in Methode B, werden durch den Parameter  $s$  die Gewichtungsfaktoren ein Stück weit vereinheitlicht. Diese Vereinheitlichung wird dann durch den Parameter  $\alpha$  fortgesetzt, falls dieser kleiner als 2 ist.

Im Vergleich zwischen Methode D<sub>0</sub> und D fällt auf, dass die Ergebnisse von Methode D näher an denen der Methode C<sub>0</sub> liegen. Interessant ist, dass durch die Parameter  $\alpha$  und  $S$  in Methode D sowohl der von Methode D<sub>0</sub> eingeschlagene Verlauf fortgeführt (vgl. Teilnehmer 1 bis 3), als auch umgekehrt (vgl. Teilnehmer 4 und 5) werden kann. Die Richtung ist in diesem Beispiel abhängig von der Lage der effektiven Teilnehmerunsicherheit aus Methode D<sub>0</sub> zu denen aus Methode C<sub>0</sub>. Methode D ändert die effektive Teilnehmerunsicherheit von Methode D<sub>0</sub> jeweils in Richtung derer aus Methode C<sub>0</sub>. Da im gezeigten Beispiel der Parameter  $S$  durch die empirische Stichprobenstandardabweichung gegeben ist, gilt hier  $S = u_{i;\text{eff};C_0} \quad \forall i \in [1; N_r]$ .

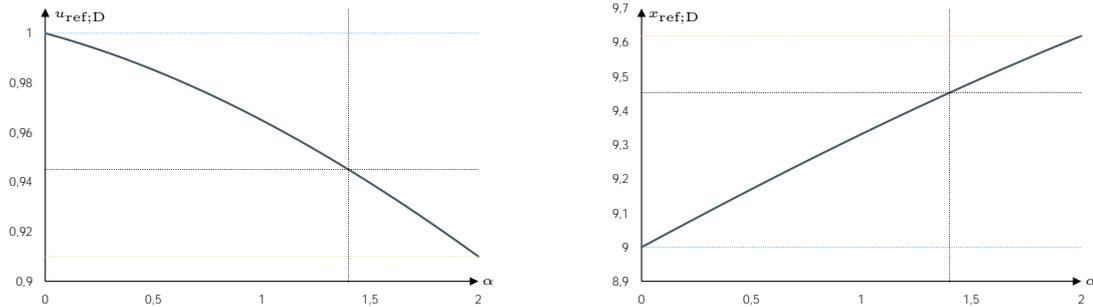


**Abbildung 6.14**

Links sind die finalen Gewichtungsfaktoren aller Methoden gezeigt. Die Methoden D<sub>0</sub> und D stellen einen Kompromiss aus individueller Gewichtung (Methode B) und einheitlicher Gewichtung (Methode C<sub>0</sub>/C) dar. Rechts ist die effektive Teilnehmerunsicherheit aus Gleichung (6.36) gezeigt. Im gewählten Beispiel entspricht der Parameter  $S$  der empirischen Stichprobenstandardabweichung, d.h. der effektiven Teilnehmerunsicherheit in Methode C<sub>0</sub>, dargestellt durch die schwarz-gepunktete Linie. Die hellgrau-gepunktete Linie zeigt informativ die Unsicherheit des Mandel-Paule Mittelwertes gleichmäßig rückgerechnet auf alle Teilnehmer, d.h.  $\sqrt{N_r}u_{\text{ref};D_0}$ . Im Vergleich zwischen Methode D<sub>0</sub> und Methode D ist letztere in beiden Abbildungen näher an an den Werten des arithmetischen Mittelwertes, da durch die Parameter  $\alpha$  und  $S$  die Vereinheitlichung des Mandel-Paule Mittelwertes fortgesetzt wird.

Falls der Parameter  $S$  durch die Unsicherheit des Mandel-Paule Mittelwertes gegeben ist, kann dieser

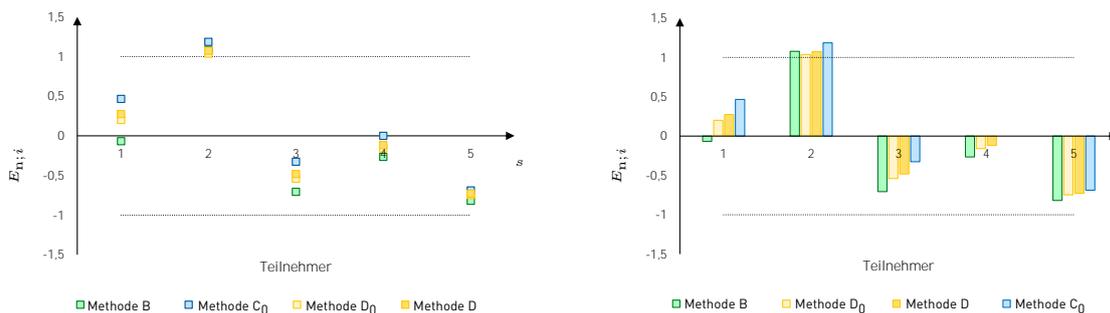
Zusammenhang nicht allgemein aufgestellt werden, d.h. dann gilt  $\exists i \in [1; N_r] S \neq u_{i;eff;D_0}$ . Der Parameter  $S$  entspricht in diesem Fall zwar einer einheitlichen Teilnehmerunsicherheit, die durch ein gleichmäßiges Rückrechnen und Aufteilen der Unsicherheit des Mandel-Paule Mittelwertes auf die einzelnen Teilnehmer entsteht, allerdings ist das nicht zu verwechseln mit der effektiven Teilnehmerunsicherheit, die nach wie vor für jeden Teilnehmer individuell ist, wenn auch im Vergleich zu Methode B vereinheitlicht. Gezeigt ist dies durch die hellgrau-gepunktete Linie in Abbildung 6.14.



**Abbildung 6.15**

Gezeigt ist der Verlauf der Referenzunsicherheit (links) und des Referenzwertes (rechts) in Abhängigkeit vom Parameter  $\alpha$ . Für  $\alpha = 2$  wird der Mandel-Paule Mittelwert aus Methode  $D_0$  und für  $\alpha = 0$  der arithmetische Wert von Methode  $C_0$  reproduziert. Im Beispiel ist  $\alpha = 1,4$ , sodass eine Zwischenform für die Referenzdaten in Methode D resultiert.

Das bereits beschriebene Verhalten der Gewichtungsfaktoren überträgt sich auch auf die Referenzdaten, wie in Abbildung 6.12 gezeigt. Der Verlauf in Abhängigkeit vom Parameter  $\alpha$  ist in Abbildung 6.15 dargestellt. Wie in Abschnitt 6.2.1 diskutiert, resultiert für  $\alpha = 2$  der Mandel-Paule Mittelwert, d.h. Methode  $D_0$  und für  $\alpha = 0$  der arithmetische Mittelwert, d.h. Methode  $C_0$ . Da im vorliegenden Beispiel  $\alpha = 1,4$  ist, sind die Referenzdaten in Methode D letztlich durch einen Zwischenform der genannten Grenzfälle gegeben.



**Abbildung 6.16**

Gezeigt ist ein Vergleich der Leistungsbewertung in den verschiedenen Auswertungsmethoden. Die Ergebnisse von Methode D sind mit jenen aus Methode  $D_0$  vergleichbar, jedoch in allen Fällen näher an Methode  $C_0$ .

Abschließend ist in Abbildung 6.16 für den betrachteten Beispieldatensatz der Vergleich der Leistungsbewertung in den verschiedenen Auswertungsmethoden gezeigt, wobei die Ergebnisse aus Abbildung

6.8 aufgegriffen werden. Es zeigt sich, dass für das vorliegenden Beispiel die Leistungsbewertung der Methoden D und  $D_0$  vergleichbar sind, erstere jedoch in allen Fällen etwas näher an Methode  $C_0$  orientiert wird. Grund dafür ist gemäß den vorhergehenden Ausführungen, dass durch den Parameter  $\alpha < 2$  die von Methode  $D_0$  begonnene Vereinheitlichung fortgesetzt wird.

Anmerkung: Durch die drei Parameter  $s$ ,  $S$  und  $\alpha$  ist das Verhalten der Referenzdaten in Methode D stark vom vorliegenden Datensatz abhängig. Die hier gezeigten Eigenschaften sind beispielhaft zu verstehen. Weitere Beispiele und Vergleiche mit den anderen Auswertungsmethoden dieses Dokumentes sind in Kapitel 7 zu finden.

## 7 Vergleich der Auswertungsmethoden

Die in den Kapiteln 4, 5 und 6 detailliert vorgestellten und charakterisierten Auswertungsmethoden B, C<sub>0</sub>/C und D<sub>0</sub>/D werden in diesem Kapitel anhand ausgewählter Beispieldaten gegenübergestellt und verglichen. Der Fokus wird dabei sowohl auf die Referenzdaten, bestehend aus Referenzwert und Referenzunsicherheit, als auch auf die die Leistungsbewertung gelegt. Für eine detaillierte Analyse der jeweiligen Methoden wird auf die entsprechenden Kapitel verwiesen. Die Ergebnisse der statistischen Auswertungsmethoden B, C<sub>0</sub>/C und D<sub>0</sub>/D sind komplex und hängen stets vom konkret vorliegenden Datensatz ab. Die gezeigten Beispiele erheben daher keinen Anspruch auf Vollständigkeit.

Für die nachfolgend betrachteten Datensätze beschreibt  $\mathbf{x}$  die Menge der Teilnehmerwerte, d.h.

$$\mathbf{x} = \{x_1; x_2; \dots; x_N\}, \quad (7.1)$$

wobei  $x_1$  der Wert des Teilnehmers 1 ist, etc. Die Unsicherheiten sind gegeben durch

$$u(\mathbf{x}) = \{u_1; u_2; \dots; u_N\}, \quad (7.2)$$

mit der zu Wert  $x_1$  von Teilnehmer 1 gehörenden Unsicherheit  $u_1$ , etc. Die Gesamtzahl der Teilnehmer ist durch  $N$  gegeben, die Anzahl der Daten, die in die Bestimmung der Referenzdaten einfließen durch  $N_r \leq N$ .

Folgende Darstellung wird verwendet:

Methode	Referenzwert	Referenzunsicherheit
B	Gleichung (4.5)	Gleichung (4.3)
C <sub>0</sub>	Gleichung (5.1)	Gleichung (5.6)
C <sub>0</sub>	Gleichung (5.1)	$\min(u_{\text{ref};C_0;1}; u_{\text{ref};C_0;2})$ , Gleichungen (5.4) und (5.5)
C	Gleichung (5.20)	Gleichung (5.22)
C	Gleichung (5.20)	$\min(u_{\text{ref};C;1}; u_{\text{ref};C;2})$ , Gleichungen (5.21) und (5.30)
D <sub>0</sub>	Gleichung (6.3)	Gleichung (6.4)
D	Gleichung (6.22)	Gleichung (6.23)

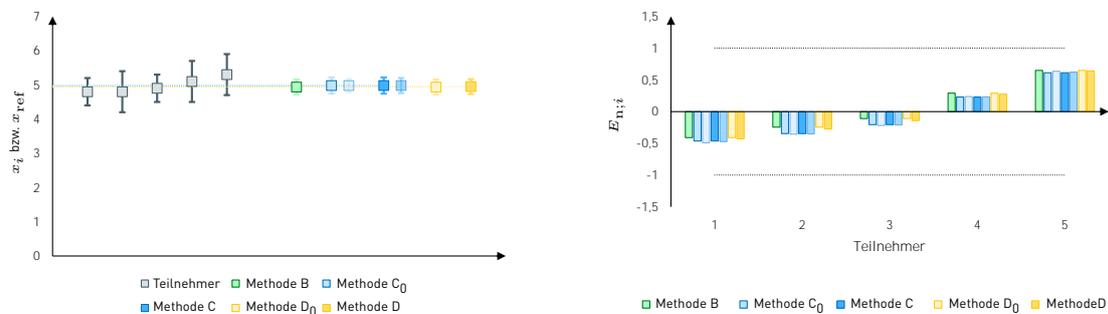
Die nachfolgenden Beispiel sind jeweils nur kurz skizziert. Eine ausführliche Analyse kann anhand der in den Kapiteln 4, 5 und 6 beschriebenen Eigenschaften durchgeführt werden.

## 7.1 Beispiel 1: Konsistenter Datensatz

Betrachtet wird der Datensatz

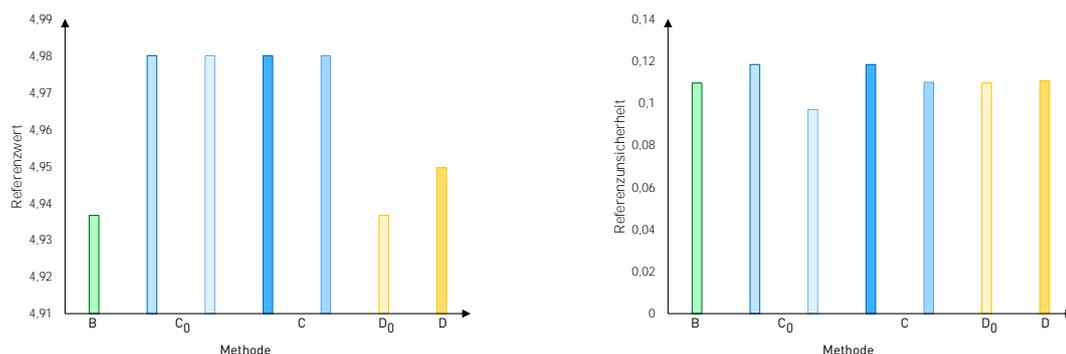
$$\begin{aligned}
 N &= 5, \\
 \mathbf{x}_1 &= \{4,8; 4,8; 4,9; 5,1; 5,3\}, \\
 u(\mathbf{x}_1) &= \{0,20; 0,30; 0,20; 0,30; 0,30\}.
 \end{aligned}
 \tag{7.3}$$

Es liegt ein konsistenter Datensatz vor, d.h. die Streuung der Messwerte ist im Rahmen der Messunsicherheiten plausibel, wie Abbildung 7.1 zeigt. Die betrachteten Methoden B, C<sub>0</sub>, C, D<sub>0</sub> und D stimmen gut überein und liefern sowohl ähnliche Referenzwerte, als auch ähnliche Referenzunsicherheiten, vgl. Abbildung 7.2. Dies zeigt sich auch in der Leistungsbewertung der Teilnehmer in Abbildung 7.1. Die  $E_{n,i}$ -Werte sind in allen Methoden in einer ähnlichen Größenordnung.



**Abbildung 7.1**

Links: Datensatz der Teilnehmer im Beispiel 1, sowie Referenzdaten in den verschiedenen Auswertungsmethoden. Rechts: Dazugehörige Leistungsbewertung der Teilnehmer mittels  $E_{n,i}$ -Wert.



**Abbildung 7.2**

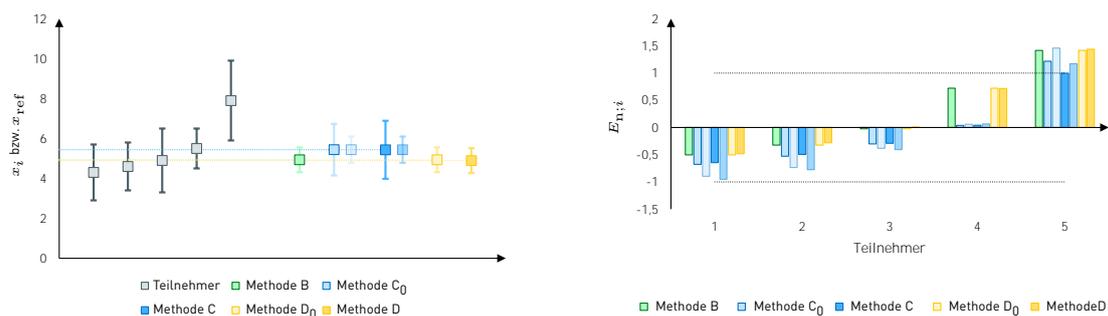
Links: Referenzwerte in den verschiedenen Auswertungsmethoden für den Datensatz aus Beispiel 1. Rechts: Dazugehörige Referenzunsicherheiten.

## 7.2 Beispiel 2: Datensatz mit Ausreißern

Betrachtet wird der Datensatz

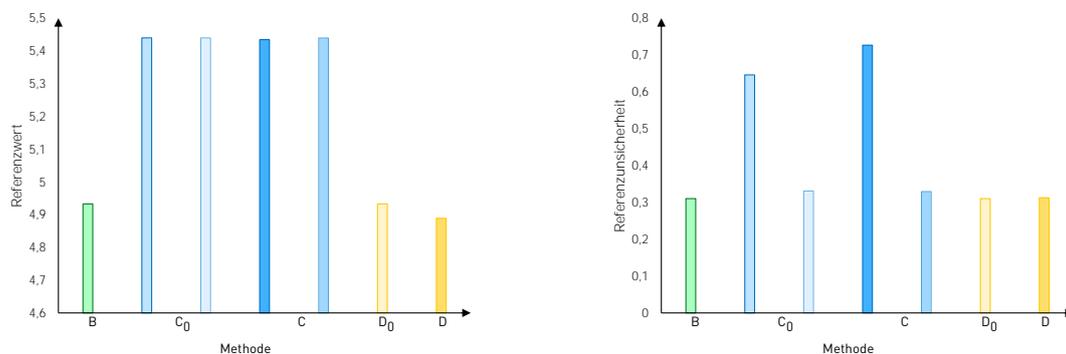
$$\begin{aligned}
 N &= 5, \\
 \mathbf{x}_2 &= \{4,3; 4,6; 4,9; 5,5; 7,9\}, \\
 u(\mathbf{x}_2) &= \{0,70; 0,60; 0,80; 0,50; 1,0\}.
 \end{aligned}
 \tag{7.4}$$

Im vorliegenden Datensatz handelt es sich bei Teilnehmer 5 um einen Ausreißer, der von den Methoden B, C, D<sub>0</sub> und D erkannt wird. Die Daten der Teilnehmer 1 bis 4 hingegen sind konsistent, wie Abbildung 7.3 zeigt. Die Methoden B, D<sub>0</sub> und D liefern daher ähnlich Ergebnisse, da der Ausreißer aus der Bestimmung der Referenzdaten entfernt wird. Methode C hingegen modifiziert den Ausreißer-Wert, sodass sich die qualitative Lage des Ausreißer in den Referenzdaten niederschlägt und Methode C in diesem Fall ähnliche Referenzdaten liefert wie Methode C<sub>0</sub>, vgl. Abbildung 7.4. Die Unterschiede in den Referenzdaten wirken sich auch auf die Leistungsbewertung aus. Signifikante Unterschiede zeigen sich zwischen den Methoden B, D<sub>0</sub> und D und den Methoden C<sub>0</sub> und C. Insbesondere für Teilnehmer 5, dessen  $E_{n;i=5}$ -Wert um  $\pm 1$  liegt, kann die Wahl der Methode über bestehen und nicht bestehen entscheiden.



**Abbildung 7.3**

Links: Datensatz der Teilnehmer im Beispiel 2, sowie Referenzdaten in den verschiedenen Auswertungsmethoden. Rechts: Dazugehörige Leistungsbewertung der Teilnehmer mittels  $E_{n;i}$ -Wert.



**Abbildung 7.4**

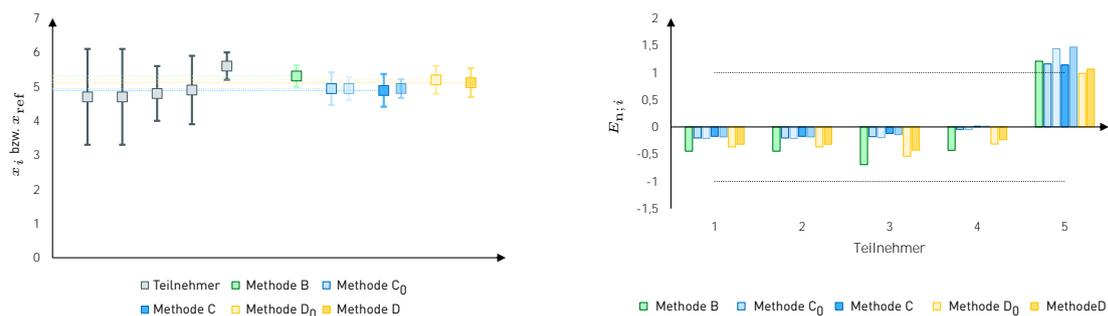
Links: Referenzwerte in den verschiedenen Auswertungsmethoden für den Datensatz aus Beispiel 2. Rechts: Dazugehörige Referenzunsicherheiten.

### 7.3 Beispiel 3: Ein Teilnehmer mit kleiner Unsicherheit

Betrachtet wird der Datensatz

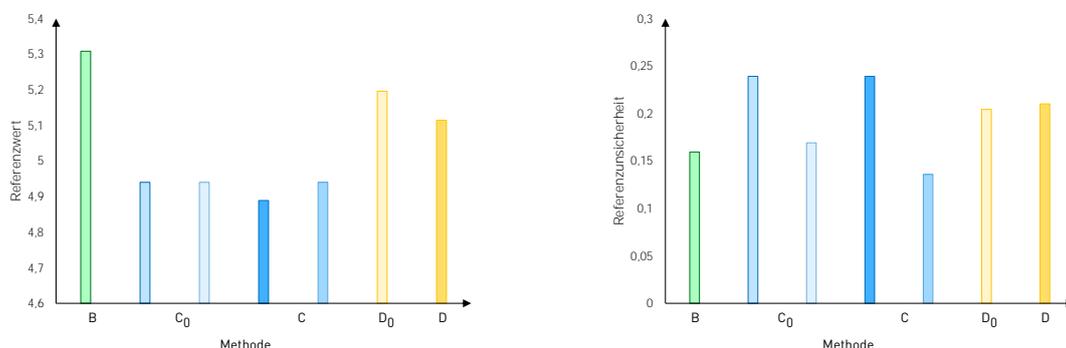
$$\begin{aligned}
 N &= 5, \\
 \mathbf{x}_3 &= \{4,7; 4,7; 4,8; 4,9; 5,6\}, \\
 u(\mathbf{x}_3) &= \{0,70; 0,70; 0,40; 0,50; 0,20\}.
 \end{aligned}
 \tag{7.5}$$

Teilnehmer 5 berichtet eine deutlich kleinere Unsicherheit als die restlichen Teilnehmer 1 bis 4 und weicht auch im Wert  $x_5$  von den restlichen Teilnehmern ab. Als Ausreißer erachtet wird der Wert jedoch nur von Methode C. Methode B ist erwartungsgemäß maßgeblich vom Teilnehmer mit der kleinen Unsicherheit beeinflusst. Die Methoden  $D_0$  und D erweisen sich als Kompromiss zwischen den Methoden B und den Methoden  $C_0/C$ , sowohl in den Referenzdaten, als auch in der Leistungsbewertung, wie die Abbildungen 7.5 und 7.6 zeigen. Signifikante Unterschiede ergeben sich für Teilnehmer 5, da hier die Wahl der Methode über eine positive oder negative Leistungsbewertung entscheiden kann. Hier zeigt sich, wie bereits in Beispiel 2, dass die Unterschiede in den verschiedenen Auswertungsmethoden vor allem dann von Bedeutung sind, wenn ein Teilnehmer einen  $E_{n;i}$ -Wert um  $\pm 1$  erzielt.



**Abbildung 7.5**

Links: Datensatz der Teilnehmer im Beispiel 3, sowie Referenzdaten in den verschiedenen Auswertungsmethoden. Rechts: Dazugehörige Leistungsbewertung der Teilnehmer mittels  $E_{n;i}$ -Wert.



**Abbildung 7.6**

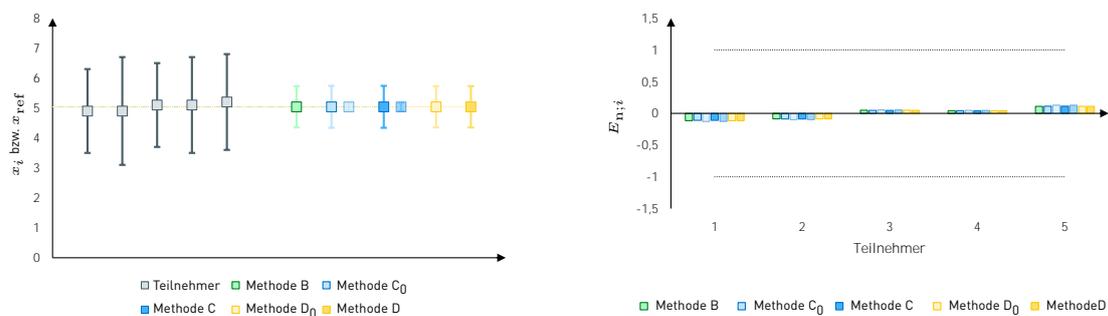
Links: Referenzwerte in den verschiedenen Auswertungsmethoden für den Datensatz aus Beispiel 3. Rechts: Dazugehörige Referenzunsicherheiten.

## 7.4 Beispiel 4: Geringe Streuung; große, ähnliche Messunsicherheit

Betrachtet wird der Datensatz

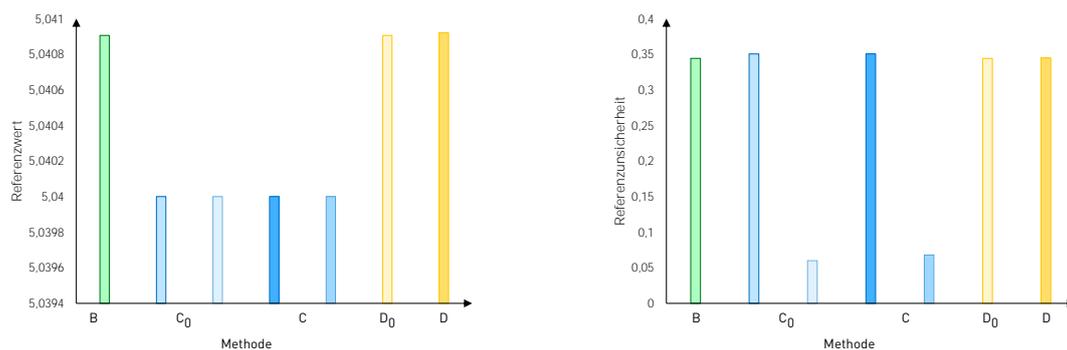
$$\begin{aligned}
 N &= 5, \\
 \mathbf{x}_4 &= \{4,9; 4,9; 5,1; 5,1; 5,2\}, \\
 u(\mathbf{x}_4) &= \{0,70; 0,90; 0,70; 0,80; 0,80\}.
 \end{aligned}
 \tag{7.6}$$

Im Vergleich zu den Teilnehmerunsicherheiten weist der Datensatz eine sehr geringe Streuung der Teilnehmerwerte auf. Wie die Abbildungen 7.7 und 7.8 zeigen, stimmen die Referenzdaten und die Leistungsbewertung in allen Methoden gut überein. Wichtig ist hier festzuhalten, dass in den Methoden C<sub>0</sub> und C die Referenzunsicherheit durch die Fehlerfortpflanzung und nicht durch die empirische Standardabweichung bestimmt ist. Letztere würde zu einer deutlichen Unterschätzung der Referenzunsicherheit führen.



**Abbildung 7.7**

Links: Datensatz der Teilnehmer im Beispiel 4, sowie Referenzdaten in den verschiedenen Auswertungsmethoden. Rechts: Dazugehörige Leistungsbewertung der Teilnehmer mittels  $E_{n,i}$ -Wert.



**Abbildung 7.8**

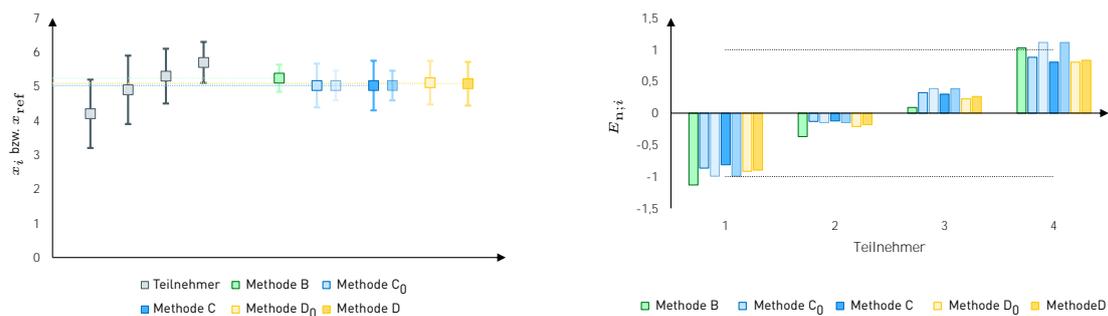
Links: Referenzwerte in den verschiedenen Auswertungsmethoden für den Datensatz aus Beispiel 4. Rechts: Dazugehörige Referenzunsicherheiten.

## 7.5 Beispiel 5: Mäßig konsistenter Datensatz

Betrachtet wird der Datensatz

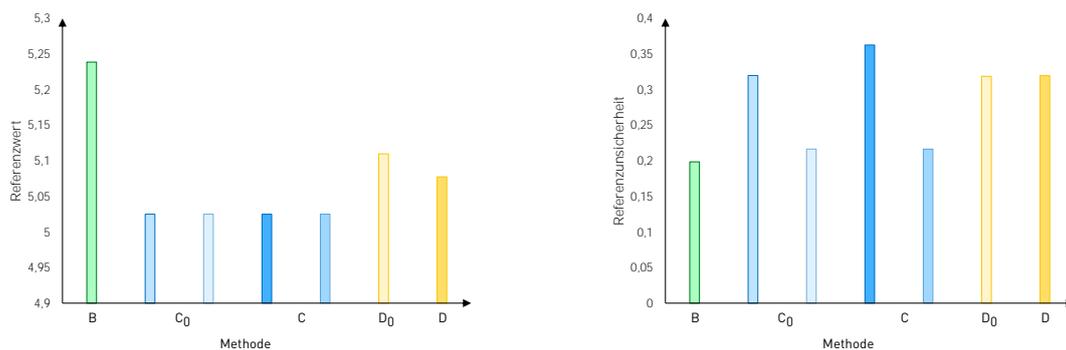
$$\begin{aligned}
 N &= 4, \\
 \mathbf{x}_5 &= \{4,2; 4,9; 5,3; 5,7\}, \\
 u(\mathbf{x}_5) &= \{0,50; 0,50; 0,40; 0,30\}.
 \end{aligned}
 \tag{7.7}$$

Für die vorliegenden mäßig konsistenten Daten erweisen sich die Methoden  $D_0$  und  $D$  als Kompromiss zwischen der unsicherheitsbasierten Methode  $B$  und den streuungsbasierten Methoden  $C_0/C$ , wie die Abbildungen 7.9 und 7.10 zeigen. Insbesondere in der Leistungsbewertung von Teilnehmer 1 und 4 entscheidet die Wahl der Methode über ein positives oder negatives bestehen des  $E_{n,i}$ -Kriteriums. Dabei liefert Methode  $B$  für die Teilnehmer 1 und 4 die größten  $E_{n,i}$ -Werte, was u.a. an der kleinen Referenzunsicherheit liegt, da die Streuung der Teilnehmerwerte in den Referenzdaten von Methode  $B$  nicht berücksichtigt wird.



**Abbildung 7.9**

Links: Datensatz der Teilnehmer im Beispiel 5, sowie Referenzdaten in den verschiedenen Auswertungsmethoden. Rechts: Dazugehörige Leistungsbewertung der Teilnehmer mittels  $E_{n,i}$ -Wert.



**Abbildung 7.10**

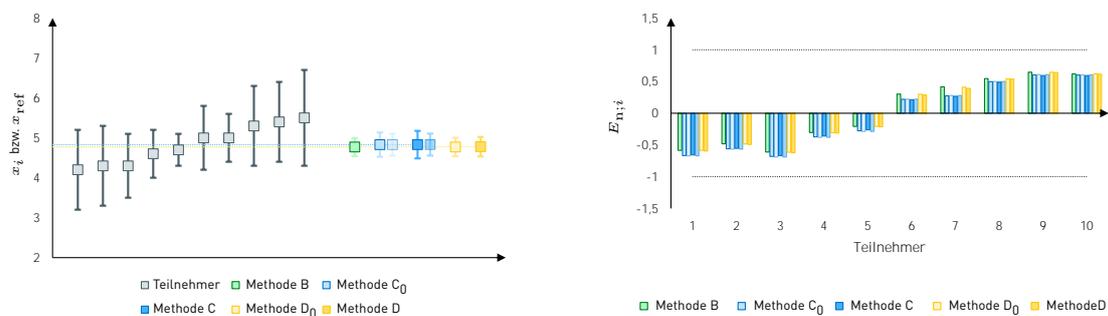
Links: Referenzwerte in den verschiedenen Auswertungsmethoden für den Datensatz aus Beispiel 5. Rechts: Dazugehörige Referenzunsicherheiten.

## 7.6 Beispiel 6: Viele Teilnehmer

Betrachtet wird der Datensatz

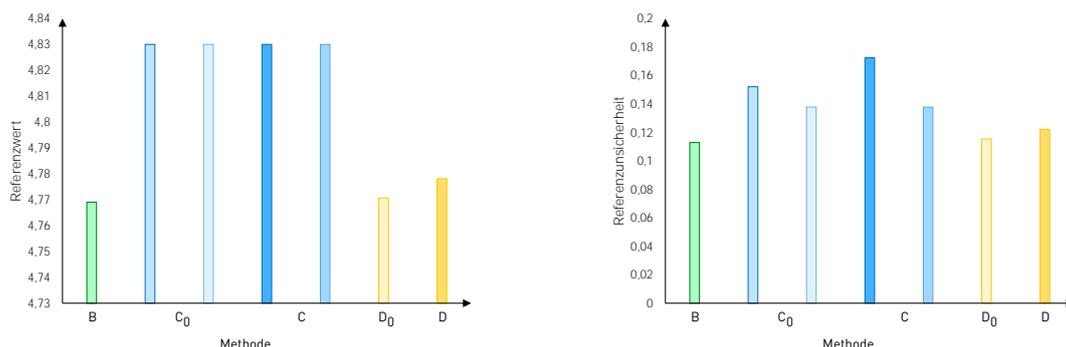
$$\begin{aligned}
 N &= 10, \\
 \mathbf{x}_6 &= \{4,2; 4,3; 4,3; 4,6; 4,7; 5,0; 5,0; 5,3; 5,4; 5,5\}, \\
 u(\mathbf{x}_6) &= \{0,50; 0,50; 0,40; 0,30; 0,20; 0,40; 0,30; 0,50; 0,50; 0,60\}.
 \end{aligned}
 \tag{7.8}$$

Bei einer Vielzahl an vorliegenden, übereinstimmenden Teilnehmerdaten zeigt sich in diesem Beispiel, dass die Ergebnisse aller Auswertungsmethoden gut übereinstimmen, sowohl im den Referenzdaten, als auch in der Leistungsbewertung, wie die Abbildungen 7.11 und 7.12 zeigen. Auffallend ist zudem, dass die Referenzunsicherheit in allen Methoden aufgrund der vielen Teilnehmer klein wird. Dies kann damit begründet werden, dass die Kenntnis des Referenzwertes umso besser wird, je mehr übereinstimmende Daten in dessen Bestimmung einfließen.



**Abbildung 7.11**

Links: Datensatz der Teilnehmer im Beispiel 6, sowie Referenzdaten in den verschiedenen Auswertungsmethoden. Rechts: Dazugehörige Leistungsbewertung der Teilnehmer mittels  $E_{n,i}$ -Wert.



**Abbildung 7.12**

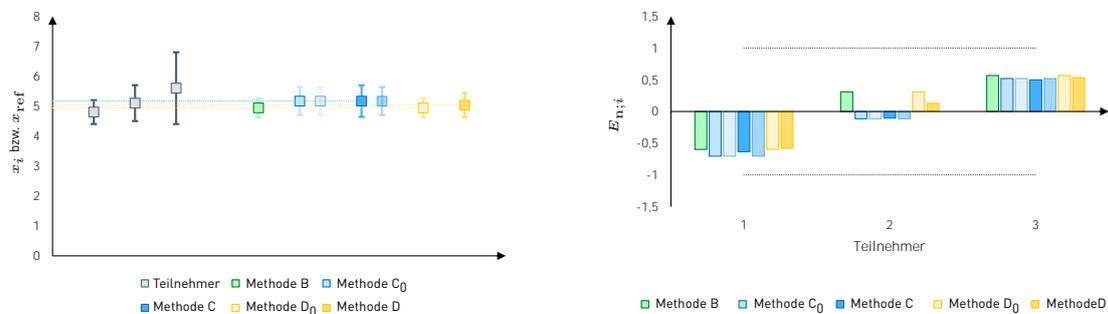
Links: Referenzwerte in den verschiedenen Auswertungsmethoden für den Datensatz aus Beispiel 6. Rechts: Dazugehörige Referenzunsicherheiten.

## 7.7 Beispiel 7: Wenig Teilnehmer

Betrachtet wird der Datensatz

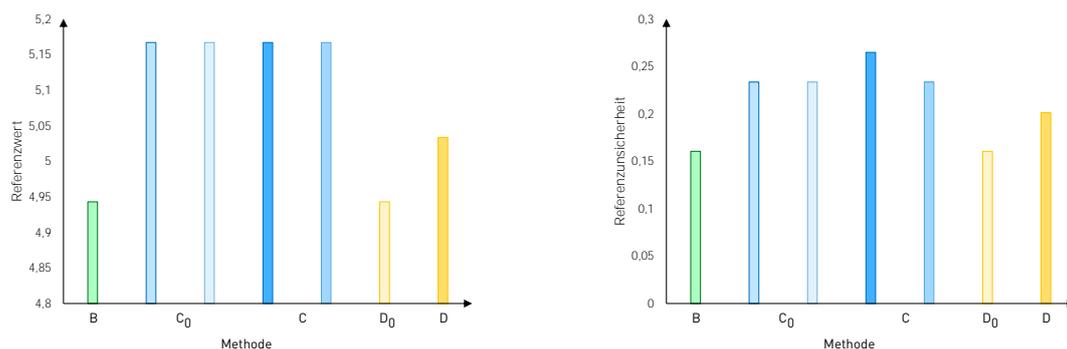
$$\begin{aligned}
 N &= 3, \\
 \mathbf{x}_7 &= \{4,8; 5,1; 5,6\}, \\
 u(\mathbf{x}_7) &= \{0,20; 0,30; 0,60\}.
 \end{aligned}
 \tag{7.9}$$

Bei vorliegen einer geringen Teilnehmerzahl zeigt sich insbesondere ein Unterschied zwischen Methode  $D_0$  und Methode D. Letztere berücksichtigt die Teilnehmerzahl über den Parameter  $\alpha$ , wodurch die Referenzdaten im Vergleich zu Methode  $D_0$  näher an Methode  $C_0/C$  liegen. Da kein Ausreißer vorliegt, sind die Ergebnisse von Methode C in den Abbildungen 7.13 und 7.14 gezeigt. Tatsächlich sollte Methode C, falls Ausreißer möglich sind, erst ab  $N = 4$  Teilnehmern angewandt werden.



**Abbildung 7.13**

Links: Datensatz der Teilnehmer im Beispiel 7, sowie Referenzdaten in den verschiedenen Auswertungsmethoden. Rechts: Dazugehörige Leistungsbewertung der Teilnehmer mittels  $E_{n,i}$ -Wert.



**Abbildung 7.14**

Links: Referenzwerte in den verschiedenen Auswertungsmethoden für den Datensatz aus Beispiel 7. Rechts: Dazugehörige Referenzunsicherheiten.

An dieser Stelle sei angemerkt, dass für die statistische Auswertung mindestens  $N_T = 3$  Datensätze vorliegen sollten, die für die Bestimmung der Referenzdaten herangezogen werden. Idealerweise liegen  $N_T = 4$  Datensätze vor, um eine sinnvolle Auswertung in allen Methoden zu gewährleisten.

## 8 Bilaterale Eignungsprüfungen

Wie in den vorhergehenden Kapiteln beschrieben, eignet sich die statistische Auswertung von Eignungsprüfungsdaten gemäß den Methoden B, C<sub>0</sub>, C, D<sub>0</sub> und D erst ab  $N_r = 3$ , bzw. besser  $N_r = 4$  Teilnehmerdaten, die in die Bestimmung der Referenzdaten einfließen. Bei vorliegen einer bilateralen Eignungsprüfung, d.h. einem Datensatz mit  $N = 2$  Teilnehmern, ist eine statistische Auswertung nicht sinnvoll und die in Kapitel 3 beschriebene Methode A ist stets auszuwählen. Begründet ist dies dadurch, dass in den B, C<sub>0</sub>, C, D<sub>0</sub> und D, das Ergebnis vorhersehbar ist, in dem Sinne, dass aufgrund analytischer Betrachtungen der Gleichungen für die Referenzdaten und die Leistungsbewertung, bereits vor der Messung klar ist, welches Ergebnis die Teilnehmer erzielen werden. In allen Methoden zeigt sich, dass bei einem bilateralen Datensatz die Leistungsbewertung immer positiv ausfällt, wenn beide Teilnehmer in die Bestimmung der Referenzdaten einfließen.

An dieser Stelle sei angemerkt, dass bei vorliegen von zwei Datensätze eine Ausreißer-Diskussion in den statistischen Methoden B, C<sub>0</sub>, C, D<sub>0</sub> und D nicht zielführend ist. Der Grund dafür ist, dass die Eliminierung eines Ausreißers dazu führt, dass ein Teilnehmer die Referenzdaten bestimmt und somit letztlich zu einem Referenzlabor gemäß Methode A würde. Zudem ist bei vorliegen zweier Datensätze nicht zwangsläufig klar, welcher Teilnehmer als Ausreißer zu erachten ist, sodass insbesondere die Methoden C<sub>0</sub>, C, D<sub>0</sub> und D keine Ausreißer identifizieren, sondern stattdessen die Streuung hinreichend groß berücksichtigt wird. Bei Methode B wird bei Fehlschlagen des  $\chi^2$ -Tests der Teilnehmer mit der größeren Unsicherheit als Ausreißer erachtet.

Der nachfolgend betrachtet Datensatz von  $N = 2$  Teilnehmern besteht aus den Teilnehmerwerten

$$\mathbf{x} = \{x_1; x_2\} \quad (8.1)$$

und den dazugehörigen Unsicherheiten

$$u(\mathbf{x}) = \{u_1; u_2\}. \quad (8.2)$$

### 8.1 Methode B

Gemäß den Gleichungen (4.5) und (4.3) sind die Referenzdaten in Methode B im Falle einer bilateralen Eignungsprüfung gegeben durch

$$u_{\text{ref};B}^2 = \frac{u_1^2 u_2^2}{u_1^2 + u_2^2} \quad (8.3)$$

und

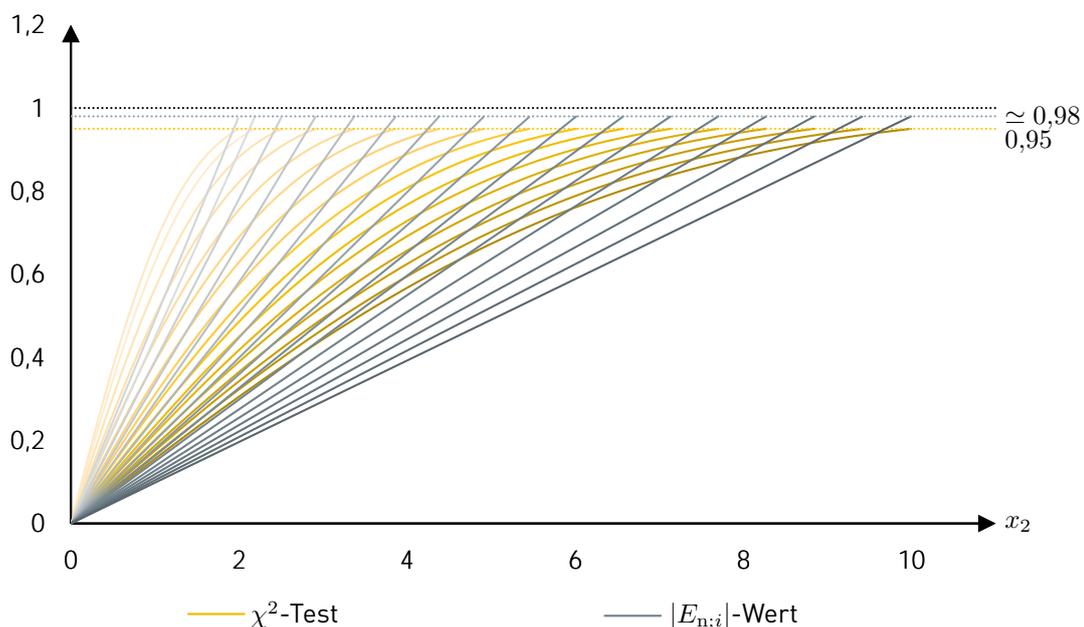
$$x_{\text{ref};B} = \frac{1}{u_1^2 + u_2^2} (u_2^2 x_1 + u_1^2 x_2). \quad (8.4)$$

Der  $\chi^2$ -Test aus Abschnitt 4.1.1 kann entweder bestanden werden, sodass  $N_r = N = 2$  ist, oder nicht bestanden werden. Ist letzteres der Fall und die berichteten Teilnehmerunsicherheiten sind

unterschiedlich, so ist der Teilnehmer mit der größeren Unsicherheit aus den Bestimmungsgleichungen der Referenzdaten zu entfernen, denn

$$\chi_{i;\text{obs}}^2 = u_i^2 \frac{(x_1 - x_2)^2}{(u_1^2 + u_2^2)^2} \quad (8.5)$$

Letztlich bedeutet das, dass der Teilnehmer mit der kleinen Unsicherheit zum Referenzlabor und Methode B zu Methode A wird. Berichten beide Teilnehmer die gleiche Unsicherheit und der  $\chi^2$ -Test schlägt fehl, so kann Methode B nicht weiter angewandt werden, da beide Teilnehmer aus der Berechnung zu nehmen sind.



**Abbildung 8.1**

Gezeigt ist der Verlauf des Betrag des  $E_{n;i}$ -Wertes (grau) und des  $\chi^2$ -Tests (gelb) für einen Datensatz mit oBdA  $x_1 = 0$  und oBdA  $u_1 = 1,0$  und variablem  $x_2$  und  $u_2$ . Letzteres nimmt Werte zwischen 0,20 (hell) und 5,0 (dunkel) in Schritten von 0,30 an. Es zeigt sich, dass der  $\chi^2$ -Test nur dann bestanden ist (d.h. kleiner 0,95), wenn beide Teilnehmer einen betragsmäßigen  $E_{n;i}$ -Wert kleiner als ca. 0,98 aufweisen und damit unabhängig von den konkreten Werten stets eine positive Leistungsbewertung erzielen.

Relevant ist daher lediglich der Fall, dass der  $\chi^2$ -Test bestanden wird, d.h. beide Teilnehmer in die Bestimmung der Referenzdaten einfließen und die Referenzdaten durch die Gleichungen (8.3) und (8.4) gegeben sind. Für den  $E_{n;i}$ -Wert der Teilnehmer gilt gemäß Gleichung (4.11) unter der Berücksichtigung der Korrelation zwischen den Teilnehmer- und Referenzdaten

$$|E_{n;i}| = \frac{|x_2 - x_1|}{2\sqrt{u_1^2 + u_2^2}} \quad (8.6)$$

Damit ist zunächst schon einmal klar, dass beide Teilnehmer den betragsmäßig gleichen  $E_{n;i}$ -Wert erzielen werden. Ein  $E_{n;i}$ -Wert größer 1 resultiert genau dann, wenn  $|x_2 - x_1| > 2\sqrt{u_1^2 + u_2^2}$ . Für den

$\chi^2$ -Test bedeutet das wiederum, dass

$$\chi_{\text{obs}}^2 = 4E_{n;i}^2 > 4, \quad (8.7)$$

was für  $N_r = 2$  zu einem Fehlschlagen des  $\chi^2$ -Tests führt und zur Nichtanwendbarkeit von Methode B. Damit ist bereits vor der Durchführung aller Messungen klar, dass Methode B entweder nicht anwendbar ist, oder beide Teilnehmer mit dem betragsmäßig gleichen  $E_{n;i}$ -Wert die Eignungsprüfung bestehen werden. Der betragsmäßig größte  $E_{n;i}$ -Wert liegt bei ca. 0,98, wie Abbildung 8.1 zeigt.

## 8.2 Methode C<sub>0</sub>

Für die Betrachtung des arithmetischen Mittelwertes sei zunächst angenommen, dass die Referenzunsicherheit durch die empirische Standardabweichung gegeben ist, d.h. durch Gleichung (5.4). In diesem Fall gilt für die Referenzdaten

$$x_{\text{ref};C_0} = \frac{x_1 + x_2}{2} \quad (8.8)$$

und

$$u_{\text{ref};C_0} = u_{\text{ref};C_0;1} = \frac{|x_1 - x_2|}{2}. \quad (8.9)$$

Für die Leistungsbewertung gilt demnach unter der Berücksichtigung der Korrelation gemäß Gleichung (5.8)

$$|E_{n;i}| = \frac{1}{2}. \quad (8.10)$$

Für den Fall, dass die Referenzunsicherheit nicht durch die empirische Standardabweichung, sondern durch die Fehlerfortpflanzung der Teilnehmerunsicherheiten gegeben ist gilt

$$u_{\text{ref};C_0} = u_{\text{ref};C_0;2} \geq u_{\text{ref};C_0;1} = \frac{|x_1 - x_2|}{2}, \quad (8.11)$$

sodass

$$|E_{n;i}| = \frac{|x_1 - x_2|}{4u_{\text{ref};C_0}} \leq \frac{1}{2}. \quad (8.12)$$

In beiden Fällen ist somit bereits vor der Durchführung aller Messungen klar, dass beide Teilnehmer eine positive Leistungsbewertung erzielen werden, mit dem betragsmäßig gleichen  $E_{n;i}$ -Wert kleiner oder gleich 0,5.

## 8.3 Methode C

Der Algorithmus von Methode C ist erst ab  $N_r = 4$  Teilnehmerdaten anwendbar. Da, wie bereits zu Beginn erwähnt, eine Ausreißerdiskussion bei zwei Datensätzen nicht zielführend ist, kann Methode C qualitativ auf Methode C<sub>0</sub> zurück geführt werden, wodurch unterstrichen wird, dass Methode C bei Vorliegen von zwei Datensätzen nicht sinnvoll angewandt werden kann.

## 8.4 Methode D<sub>0</sub>

Betrachtet wird hier der Fall  $s > 0$ , da für  $s = 0$  die Ergebnisse von Methode B aus Abschnitt 8.1 greifen. Der Parameter  $s$  ist gegeben durch

$$s^2 = \frac{(x_1 - x_2)^2 - (u_1^2 + u_2^2)}{2} \quad (8.13)$$

und nur definiert für  $(x_1 - x_2)^2 \geq u_1^2 + u_2^2$ , was an dieser Stelle vorausgesetzt ist, da ansonsten  $s = 0$  resultiert.

Die Referenzunsicherheit ist damit gegeben durch

$$u_{\text{ref};D_0}^2 = \frac{1}{4} \left( (x_1 - x_2)^2 - \frac{(u_1^2 - u_2^2)^2}{(x_1 - x_2)^2} \right) \quad (8.14)$$

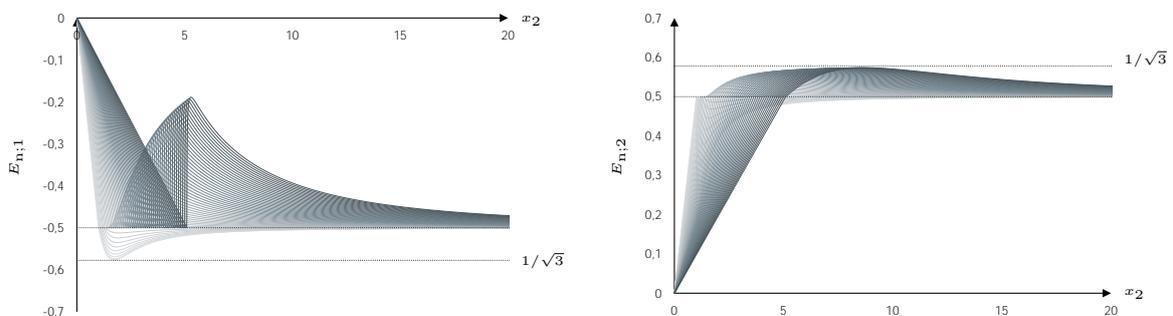
und der Referenzwert durch

$$x_{\text{ref};D_0} = \frac{1}{2} \left( (x_1 + x_2) - \frac{u_1^2 - u_2^2}{(x_1 - x_2)} \right). \quad (8.15)$$

Auf die ausformulierte Darstellung des  $E_{n;i}$ -Wertes wird an dieser Stelle verzichtet. Numerisch zeigt sich jedoch, wie in Abbildung 8.2 dargestellt, dass

$$|E_{n;i}| \leq \frac{1}{\sqrt{3}} \simeq 0,58 \quad \forall i \in [1; 2]. \quad (8.16)$$

Damit ist auch in Methode D<sub>0</sub> bereits vor der Durchführung aller Messungen klar, dass beide Teilnehmer die Eignungsprüfung bestehen werden, allerdings möglicherweise mit unterschiedlichen  $E_{n;i}$ -Werten.



**Abbildung 8.2**

Gezeigt ist der Verlauf des  $E_{n;i}$ -Wertes für einen Datensatz mit oBdA  $x_1 = 0$  und oBdA  $u_1 = 1,0$  und variablem  $x_2$  und  $u_2$ . Letzteres nimmt Werte zwischen 0,10 (hellgrau) und 5,0 (dunkelgrau) in Schritten von 0,10 an. Es zeigt sich, dass beide Teilnehmer einen betragsmäßigen  $E_{n;i}$ -Wert kleiner als  $1/\sqrt{3}$  aufweisen und damit unabhängig von den konkreten Werten stets eine positive Leistungsbewertung erzielen.

Da der  $E_{n;i}$ -Wert der beiden Teilnehmer stets kleiner als 1 ist, ist das Ausschlusskriterium aus Abschnitt 6.2.2 stets erfüllt und beide Teilnehmer fließen in die Bestimmung der Referenzdaten ein, unabhängig von den konkreten Messwerten.

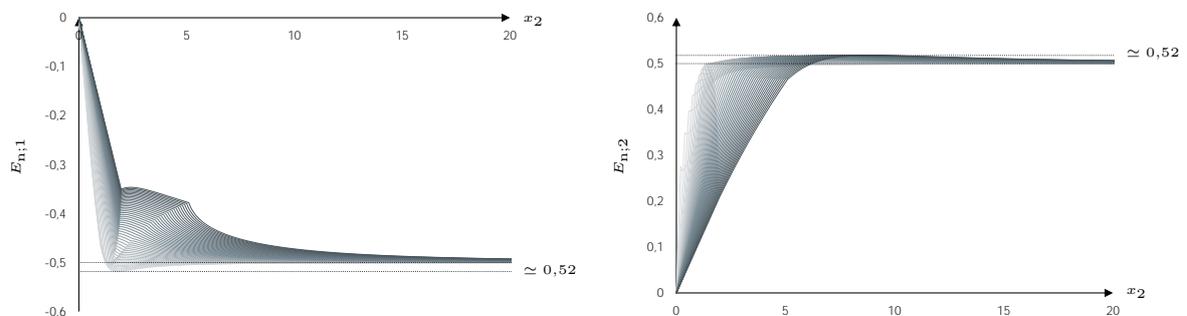
## 8.5 Methode D

Wie in Kapitel 6 gezeigt moderiert Methode D zwischen Methode C<sub>0</sub> und D<sub>0</sub> und modifiziert die Werte von Methode D<sub>0</sub> in Richtung der Werte von Methode C<sub>0</sub>.

Für Methode C<sub>0</sub> in Abschnitt 8.2 wurde gezeigt, dass der betragsmäßige  $E_{n;i}$ -Wert für beide Teilnehmer kleiner oder gleich 0,5 gegeben ist und für Methode D<sub>0</sub>, dass der betragsmäßige  $E_{n;i}$ -Wert kleiner oder gleich  $1/\sqrt{3}$  ist. Somit folgt für den maximalen  $E_{n;i}$ -Wert von Methode D, dass dieser zwischen 0,5 und  $1/\sqrt{3}$  zu finden ist. Entsprechend Abbildung 8.3 zeigt sich numerisch, dass

$$|E_{n;i}| \lesssim 0,52 \quad \forall i \in [1; 2]. \quad (8.17)$$

Somit zeigt sich auf für Methode D, dass das Ergebnis bereits vor der Durchführung aller Messungen feststeht und beide Teilnehmer eine positive Leistungsbewertung erzielen. Auf eine ausführliche Darstellung der Berechnungsgleichungen wird an dieser Stelle verzichtet.



**Abbildung 8.3**

Gezeigt ist der Verlauf des  $E_{n;i}$ -Wertes für einen Datensatz mit oBdA  $x_1 = 0$  und oBdA  $u_1 = 1,0$  und variablem  $x_2$  und  $u_2$ . Letzteres nimmt Werte zwischen 0,10 (hellgrau) und 5,0 (dunkelgrau) in Schritten von 0,10 an. Es zeigt sich, dass beide Teilnehmer einen betragsmäßigen  $E_{n;i}$ -Wert kleiner als ca. 0,52 aufweisen und damit unabhängig von den konkreten Werten stets eine positive Leistungsbewertung erzielen.

## Abbildungsverzeichnis

2.1	Normalverteilung eines Teilnehmerergebnisses $x_i \pm U_i$ . . . . .	3
2.2	Zusammenhang zwischen Überlappung der Unsicherheitsbalken und einer positiven $E_{n;i}$ -Bewertung . . . . .	5
2.3	Positive $E_{n;i}$ -Bewertung bei Werten innerhalb des Unsicherheitsintervall . . . . .	6
2.4	Leistungsbewertung für $u_i < u_{\text{ref}}$ . . . . .	8
2.5	Normalverteilung der statistischen Größe $\mathcal{E}_{n;i}$ . . . . .	9
2.6	Annahme und Ablehnungsbereich der Nullhypothese . . . . .	11
2.7	$\beta$ -Fehler des Hypothesentests bei konkretisierter Alternativhypothese . . . . .	12
4.1	Methode B für zwei Teilnehmer mit gleicher Unsicherheit . . . . .	26
4.2	Methode B für zwei Teilnehmer mit $u_1 \ll u_2$ . . . . .	27
4.3	Methode B für einen Ausreißer mit kleiner Unsicherheit . . . . .	28
4.4	Methode B $E_{n;i}$ -Wert für einen Ausreißer mit kleiner Unsicherheit . . . . .	30
4.5	Methode B beispielhafte Abhängigkeit von der Teilnehmerzahl . . . . .	31
5.1	Methode C <sub>0</sub> Streuung vs. Teilnehmerunsicherheiten . . . . .	36
5.2	Methode C <sub>0</sub> Korrelationskoeffizienten und $E_{n;i}$ -Wert . . . . .	37
5.3	Methode C <sub>0</sub> vs. Methode B Korrelationskoeffizienten und $E_{n;i}$ -Wert . . . . .	38
5.4	Methode C <sub>0</sub> vs. Methode B Verlauf der Referenzunsicherheit in Abhängigkeit einer Teilnehmerunsicherheit . . . . .	39
5.5	Methode C <sub>0</sub> Ausreißereinfluss . . . . .	40
5.6	Arithmetischer Mittelwert vs. Median . . . . .	41
5.7	Methode C Beispieldatensatz Startwerte . . . . .	42
5.8	Methode C Beispieldatensatz Datenmodifikation . . . . .	43
5.9	Methode C Beispieldatensatz Konvergenzverhalten . . . . .	44
5.10	Methode C modifizierte Verteilung und Korrekturfaktor . . . . .	45
5.11	Methode C Abhängigkeit des Referenzwertes von einem Teilnehmerwert . . . . .	47
5.12	Methode C vs. arithmetische Mittelwerte bzgl. der Abhängigkeit des Referenzwertes von einem Teilnehmerwert bei bis zu einem Ausreißer . . . . .	50
5.13	Methode C vs. arithmetische Mittelwerte bzgl. der Abhängigkeit des Referenzwertes von einem Teilnehmerwert bei bis zu zwei Ausreißern . . . . .	51
5.14	Methode C vs. arithmetische Mittelwerte bzgl. der Abhängigkeit der Referenzunsicherheit von einem Teilnehmerwert bei bis zu einem Ausreißer . . . . .	52
5.15	Methode C vs. arithmetische Mittelwerte bzgl. der Abhängigkeit der Referenzunsicherheit von einem Teilnehmerwert bei bis zu zwei Ausreißern . . . . .	53
5.16	Methode C Abhängigkeit der Referenzunsicherheit von einem Teilnehmerwert mit einem festen Ausreißer nach unten vs. nach oben . . . . .	55
5.17	Methode C Leistungsbewertung eines Teilnehmers bei maximal einem Ausreißer . . . . .	55
5.18	Methode C Leistungsbewertung eines Teilnehmers bei mindestens einem Ausreißer nach unten bzw. nach oben . . . . .	56
6.1	Methode D <sub>0</sub> Verlauf der Hilfsgröße $\tilde{\chi}_{\text{obs}}^2(s)$ in Abhängigkeit von der Teilnehmerunsicherheit . . . . .	60

6.2	Methode $D_0$ Verlauf des Parameters $s$ und der Größe $\sqrt{u^2 + s^2}$ in Abhängigkeit von der Unsicherheit $u$ . . . . .	61
6.3	Methode $D_0$ Beispieldatensatz mit Referenzdaten der Methoden B, $C_0$ , C und $D_0$ . . . . .	65
6.4	Methode $D_0$ Verlauf der Gewichtungsfaktoren in Abhängigkeit vom Parameter $s$ für den Beispieldatensatz . . . . .	66
6.5	Methode $D_0$ Verlauf der Variable $\chi_{\text{obs}}^2$ zur Bestimmung des Parameters $s$ , sowie dessen Auswirkung auf die Teilnehmerunsicherheiten . . . . .	67
6.6	Methode $D_0$ Verlauf des Referenzwertes und der Referenzunsicherheit in Abhängigkeit vom Parameter $s$ für den Beispieldatensatz . . . . .	68
6.7	Methode $D_0$ Verlauf der Kovarianz und des Korrelationskoeffizient in Abhängigkeit vom Parameter $s$ für den Beispieldatensatz . . . . .	68
6.8	Methode $D_0$ Verlauf Leistungsbewertung in Abhängigkeit vom Parameter $s$ für den Beispieldatensatz . . . . .	69
6.9	Methode D Wertebereich des Parameters $\alpha$ . . . . .	71
6.10	Methode D schematische Darstellung der durch die Parameter $\alpha$ und $s$ realisierbaren Grenzfälle . . . . .	74
6.11	Methode D Beispieldatensatz mit Referenzdaten der Methoden B, $C_0$ , C, $D_0$ und D . . . . .	76
6.12	Vergleich der Referenzdaten in den verschiedenen Auswertungsmethoden am Datenbeispiel aus Gleichung [6.18] . . . . .	77
6.13	Methode D Verlauf der Gewichtungsfaktoren und deren Komponenten . . . . .	78
6.14	Vergleich der finalen Gewichtungsfaktoren und der effektiven Teilnehmerunsicherheit aller Methoden . . . . .	78
6.15	Methode D Verlauf der Referenzdaten in Abhängigkeit vom Parameter $\alpha$ . . . . .	79
6.16	Methode D Vergleich der Leistungsbewertung in den verschiedenen Auswertungsmethoden für den Beispieldatensatz . . . . .	79
7.1	Beispiel 1: Datensatz und $E_{n;i}$ -Werte . . . . .	82
7.2	Beispiel 1: Referenzwerte und Referenzunsicherheiten . . . . .	82
7.3	Beispiel 2: Datensatz und $E_{n;i}$ -Werte . . . . .	83
7.4	Beispiel 2: Referenzwerte und Referenzunsicherheiten . . . . .	83
7.5	Beispiel 3: Datensatz und $E_{n;i}$ -Werte . . . . .	84
7.6	Beispiel 3: Referenzwerte und Referenzunsicherheiten . . . . .	84
7.7	Beispiel 4: Datensatz und $E_{n;i}$ -Werte . . . . .	85
7.8	Beispiel 4: Referenzwerte und Referenzunsicherheiten . . . . .	85
7.9	Beispiel 5: Datensatz und $E_{n;i}$ -Werte . . . . .	86
7.10	Beispiel 5: Referenzwerte und Referenzunsicherheiten . . . . .	86
7.11	Beispiel 6: Datensatz und $E_{n;i}$ -Werte . . . . .	87
7.12	Beispiel 6: Referenzwerte und Referenzunsicherheiten . . . . .	87
7.13	Beispiel 7: Datensatz und $E_{n;i}$ -Werte . . . . .	88
7.14	Beispiel 7: Referenzwerte und Referenzunsicherheiten . . . . .	88
8.1	Bilaterale Eignungsprüfung Verhalten des $E_{n;i}$ -Wertes in Methode B . . . . .	90
8.2	Bilaterale Eignungsprüfung Verhalten des $E_{n;i}$ -Wertes in Methode $D_0$ . . . . .	92
8.3	Bilaterale Eignungsprüfung Verhalten des $E_{n;i}$ -Wertes in Methode D . . . . .	93

## Literaturverzeichnis

- [1] Deutsches Institut für Normung / International Organization for Standardization. *Allgemeine Anforderungen an die Kompetenz von Prüf- und Kalibrierlaboratorien*. DIN EN ISO/IEC 17025:2018-03. 2018. URL: <https://www.beuth.de/de/norm/din-en-iso-iec-17025/278030106>.
- [2] Deutsches Institut für Normung / International Organization for Standardization. *Konformitätsbewertung — Allgemeine Anforderungen an Eignungsprüfungen*. DIN EN ISO/IEC 17043:2023-10. 2023. URL: <https://www.beuth.de/de/norm/din-en-iso-iec-17043/369196889>.
- [3] International Organization for Standardization. *Statistical methods for use in proficiency testing by interlaboratory comparison*. ISO 13528:2022. 2022. URL: <https://www.iso.org/standard/78879.html>.
- [4] Maurice Cox. „Evaluation of key comparison data“. In: *Metrologia* 39 (Feb. 2003), S. 589–595. DOI: 10.1088/0026-1394/39/6/10.
- [5] John Mandel und Robert C. Paule. „Interlaboratory evaluation of a material with unequal numbers of replicates“. In: *Analytical Chemistry* 42 (1970), S. 1194–1197.
- [6] Robert C. Paule und John Mandel. „Consensus Values and Weighting Factors“. In: *Journal of research of the National Bureau of Standards* 87-5 (1982), S. 377–385.
- [7] Stefaan Pommé. *Determination of a reference value, associated standard uncertainty and degrees of equivalence*. JRC Scientific and policy reports. 2012. URL: <https://www.bipm.org/documents/20126/48196961/working-document-ID-5867/a10bc233-27a6-5259-07ed-0166066faa4d>.
- [8] Stefaan Pommé und John Keightley. „Determination of a reference value and its uncertainty through a power-moderated mean“. In: *Metrologia* 52 (Juni 2015), S200–S212. DOI: 10.1088/0026-1394/52/3/S200.
- [9] Deutsches Institut für Normung / International Organization for Standardization (Burghart Brinkmann). *Internationales Wörterbuch der Metrologie — Grundlegende und allgemeine Begriffe und zugeordnete Benennungen (VIM)*. Deutsch-englische Fassung ISO/IEC-Leitfaden 99:2007. 2007. URL: <https://www.beuth.de/de/publikation/internationales-woerterbuch-metrologie-vim-/147547918>.
- [10] BIPM. *Mutual recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes*. CIPM MRA documents. 1999 (Technical Supplement revised in 2003). URL: <https://www.bipm.org/en/cipm-mra/cipm-mra-documents>.
- [11] BIPM u. a. *Evaluation of measurement data — Guide to the expression of uncertainty in measurement*. Joint Committee for Guides in Metrology, JCGM 100:2008. 2008. URL: <https://www.bipm.org/en/committees/jc/jcgm/publications>.
- [12] Manuel Marschall und Gerd Wübbeler und Michael Borys und Clemens Elster. „On modelling of artefact instability in interlaboratory comparisons“. In: *Metrologia* 60.4 (2023), S. 045010. DOI: 10.1088/1681-7575/ace18f.
- [13] Oliver Deiser. *Die p-Normen*. Analysis 2 — Topologische Grundbegriffe. 16.01.2023, 14:17. URL: <https://www.aleph1.info/Resource?method=get&obj=Pdf&name=analysis2.pdf&pagestart=172&pageend=175>.

- [14] K. Hartmann, J. Krois und A. Rudolph. *Statistics and Geodata Analysis using R (SOGA-R)*. Department of Earth Sciences, Freie Universitaet Berlin. 2023, aufgerufen am 18.09.2024, 09:46. URL: <https://www.geo.fu-berlin.de/en/v/soga-r/Basics-of-statistics/Hypothesis-Tests/Introduction-to-Hypothesis-Testing/index.html>.
- [15] Wikipedia. *Fehler 1. und 2. Art*. 18.09.2024, 09:56. URL: [https://de.wikipedia.org/wiki/Fehler\\_1.\\_und\\_2.\\_Art](https://de.wikipedia.org/wiki/Fehler_1._und_2._Art).
- [16] Wikipedia. *Gauß-Test*. 18.09.2024, 10:06. URL: <https://de.wikipedia.org/wiki/Gau%C3%9F-Test>.
- [17] Ilja N. Bronstein u. a. *Taschenbuch der Mathematik*. Verlag Harri Deutsch, 2001.
- [18] Frank E. Grubbs. „Procedures for Detecting Outlying Observations in Samples“. In: *Technometrics* 11.1 (1969), S. 1–21. DOI: 10.1080/00401706.1969.10490657.
- [19] Frank E. Grubbs. „Sample Criteria for Testing Outlying Observations“. In: *The Annals of Mathematical Statistics* 21.1 (1950), S. 27 –58. DOI: 10.1214/aoms/1177729885.
- [20] Peter J. Huber. *Robust Statistics*. John Wiley & Sons, 1981.



## Sicherheit

braucht **Vertrauen**. Wir machen

Vertrauen **messbar**.

Für **Entscheidungen** von heute  
und **Produkte**  
der **Zukunft**.



## Kontakt



+49 8141 88887-0



pt@esz-ag.de  
www.esz-ag.de



Max-Planck-Str. 16  
82216 Eichenau