

DIN ISO 13528

ICS 03.120.30

Einsprüche bis 2018-04-08
Vorgesehen als Ersatz für
DIN ISO 13528:2009-01

Entwurf

**Statistische Verfahren für Eignungsprüfungen durch Ringversuche
(ISO 13528:2015);
Text Deutsch und Englisch**

Statistical methods for use in proficiency testing by interlaboratory comparison
(ISO 13528:2015);
Text in German and English

Méthodes statistiques utilisées dans les essais d'aptitude par comparaison interlaboratoires
(ISO 13528:2015);
Texte en allemand et anglais

Anwendungswarnvermerk

Dieser Norm-Entwurf mit Erscheinungsdatum 2017-12-08 wird der Öffentlichkeit zur Prüfung und
Stellungnahme vorgelegt.

Weil die beabsichtigte Norm von der vorliegenden Fassung abweichen kann, ist die Anwendung dieses Entwurfs
besonders zu vereinbaren.

Stellungnahmen werden erbeten

- vorzugsweise online im Norm-Entwurfs-Portal von DIN unter www.din.de/go/entwuerfe bzw. für Norm-Entwürfe der DKE auch im Norm-Entwurfs-Portal der DKE unter www.entwuerfe.normenbibliothek.de, sofern dort wiedergegeben;
- oder als Datei per E-Mail an naw@din.de möglichst in Form einer Tabelle. Die Vorlage dieser Tabelle kann im Internet unter www.din.de/go/stellungnahmen-norm-entwuerfe oder für Stellungnahmen zu Norm-Entwürfen der DKE unter www.dke.de/stellungnahme abgerufen werden;
- oder in Papierform an den DIN-Normenausschuss Wasserwesen (NAW), 10772 Berlin, Burggrafenstr. 6, 10787 Berlin.

Die Empfänger dieses Norm-Entwurfs werden gebeten, mit ihren Kommentaren jegliche relevanten
Patentrechte, die sie kennen, mitzuteilen und unterstützende Dokumentationen zur Verfügung zu stellen.

Gesamtumfang 209 Seiten

DIN-Normenausschuss Wasserwesen (NAW)
DIN-Normenausschuss Qualitätsmanagement, Statistik und Zertifizierungsgrundlagen (NQSZ)

Inhalt

	Seite
Nationales Vorwort.....	5
Nationaler Anhang NA (informativ) Literaturhinweise	6
Vorwort	7
0.1 Zweck von Eignungsprüfungen	9
0.2 Begründung für die Bewertung durch Scores in Eignungsprüfungsprogrammen	9
0.3 ISO 13528 und ISO/IEC 17043	9
0.4 Statistischer Sachverstand	10
0.5 Computersoftware	10
1 Anwendungsbereich.....	11
2 Normative Verweisungen	11
3 Begriffe	12
4 Allgemeine Grundsätze	15
4.1 Allgemeine Anforderungen an statistische Verfahren	15
4.2 Grundmodell	16
4.3 Allgemeine Ansätze für die Leistungsbewertung	16
5 Leitlinien für die statistische Planung von Eignungsprüfungsprogrammen.....	17
5.1 Einleitung zur statistischen Planung von Eignungsprüfungsprogrammen	17
5.2 Grundlage einer statistischen Planung.....	17
5.3 Überlegungen zur statistischen Verteilung der Ergebnisse	18
5.4 Überlegungen für kleine Teilnehmeranzahlen	19
5.5 Leitlinien zur Auswahl des Berichtsformats	20
6 Leitlinien für die anfängliche Überprüfung von Prüfgegenständen für die Eignungsprüfungen und von Ergebnissen.....	22
6.1 Homogenität und Stabilität von Prüfgegenständen für die Eignungsprüfung.....	22
6.2 Betrachtungen für verschiedene Messverfahren	23
6.3 Entfernung grober Fehler	24
6.4 Visuelle Überprüfung von Daten.....	24
6.5 Robuste statistische Verfahren	24
6.6 Ausreißertechniken für Einzelergebnisse	25
7 Bestimmung des zugewiesenen Werts und seiner Standardunsicherheit.....	26
7.1 Auswahl des Verfahrens zur Bestimmung des zugewiesenen Werts	26
7.2 Bestimmung der Unsicherheit des zugewiesenen Werts.....	27
7.3 Bekannte Werte aus der Probenherstellung.....	28
7.4 Zertifiziertes Referenzmaterial.....	29
7.5 Ergebnisse aus einem einzelnen Laboratorium.....	29
7.6 Konsenswert von Expertenlaboratorien	30
7.7 Konsenswert aus Teilnehmerergebnissen.....	31
7.8 Vergleich des zugewiesenen Werts mit einem unabhängigen Referenzwert	33
8 Bestimmung von Kriterien für die Leistungsbewertung	34
8.1 Ansätze für die Bestimmung von Bewertungskriterien.....	34
8.2 Werte aus Erkenntnissen von Experten.....	34

	Seite
8.3	Werte aus Erfahrungen aus früheren Runden eines Eignungsprüfungsprogramms..... 35
8.4	Wert aus einem allgemeinen Modell..... 35
8.5	Werte aus Wiederhol- und Vergleichstandardabweichungen eines früheren Ringversuchs zur Präzision eines Messverfahrens 36
8.6	Werte aus Daten, die in derselben Runde eines Eignungsprüfungsprogramms erzielt wurden 37
8.7	Überwachung der Übereinstimmung zwischen Laboratorien 38
9	Berechnung von Leistungskenngrößen 39
9.1	Allgemeine Erwägungen für die Leistungsbestimmung 39
9.2	Begrenzung der Unsicherheit des zugewiesenen Werts..... 39
9.3	Schätzwerte der Abweichung (Messabweichung)..... 40
9.4	z-Scores 41
9.5	z'-Scores 42
9.6	Zeta-Scores (ζ) 43
9.7	E_n -Scores 45
9.8	Bewertung der Teilnehmerunsicherheiten bei der Prüfung 45
9.9	Kombinierte Leistungs-Scores..... 47
10	Grafische Verfahren zur Beschreibung von Leistungswerten 47
10.1	Anwendung grafischer Verfahren 47
10.2	Histogramme von Ergebnissen oder Leistungswerten 48
10.3	Kerndichtediagramme 48
10.4	Säulendiagramme von standardisierten Leistungswerten..... 50
10.5	Youden-Diagramm 50
10.6	Diagramme von Wiederholstandardabweichungen 51
10.7	Geteilte Proben..... 52
10.8	Grafische Verfahren zur Kombination von Leistungswerten über mehrere Runden eines Eignungsprüfungsprogramms..... 53
11	Planung und Analyse von qualitativen Eignungsprüfungsprogrammen (einschließlich nominaler und ordinaler Eigenschaften) 54
11.1	Arten von qualitativen Daten..... 54
11.2	Statistische Planung 54
11.3	Zugewiesene Werte für qualitative Eignungsprüfungsprogramme 55
11.4	Leistungsbewertung und Bewertung durch Scores bei qualitativen Eignungsprüfungsprogrammen..... 57
Anhang A (normativ) Symbole und Abkürzungen 59	
Anhang B (normativ) Homogenität und Stabilität von Prüfgegenständen für die Eignungsprüfung 61	
B.1	Allgemeines Verfahren zur Homogenitätsprüfung..... 61
B.2	Beurteilungskriterien für eine Homogenitätsprüfung..... 62
B.3	Formeln für die Homogenitätsprüfung 64
B.4	Verfahren zur Stabilitätsprüfung 66
B.4.1	Allgemeine Betrachtungen zur Stabilitätsprüfung 66
B.4.2	Verfahren zur Stabilitätsprüfung im Verlauf einer Eignungsprüfungsrunde..... 66
B.5	Beurteilungskriterium für eine Stabilitätsprüfung..... 68
B.6	Stabilität unter Transportbedingungen..... 69
Anhang C (normativ) Robuste Analyse..... 70	
C.1	Robuste Analyse: Einleitung..... 70
C.2	Einfache ausreißerresistente Schätzfunktionen für den Mittelwert und die Standardabweichung der Grundgesamtheit 71
C.2.1	Median 71
C.2.2	Skalierter Median der absoluten Abweichungen <i>MADE</i> 71

	Seite
C.2.3 Normierter Interquartilsabstand $nIQR$	71
C.3 Robuste Analyse: Algorithmus A	72
C.3.1 Algorithmus A mit iterierter Skala	72
C.3.2 Varianten von Algorithmus A	73
C.4 Robuste Analyse: Algorithmus S	74
C.5 Rechenintensive robuste Schätzfunktionen: Q-Verfahren und Hampel-Schätzer	75
C.5.1 Begründung für rechenintensive Schätzfunktionen	75
C.5.2 Bestimmung einer robusten Standardabweichung unter Anwendung von Q- und Q_n - Verfahren	76
C.5.3 Bestimmung eines robusten Mittelwerts mit dem Hampel-Schätzer	79
C.5.4 Q/Hampel-Verfahren	81
C.6 Andere robuste Verfahren	81
Anhang D (informativ) Zusätzlicher Leitfaden zu statistischen Verfahren	82
D.1 Verfahren für kleine Teilnehmeranzahlen	82
D.1.1 Allgemeine Überlegungen	82
D.1.2 Verfahren zur Ausreißererkennung	82
D.1.3 Verfahren für lagebezogene Schätzwerte	83
D.1.4 Verfahren für Schätzwerte der Streuung	83
D.2 Effizienz und Bruchpunkte für robuste Verfahren	84
D.2.1 Verschiedene statistische Schätzer (z. B. robuste Verfahren) können anhand von drei Schlüsselmerkmalen verglichen werden:	84
D.2.2 Bruchpunkt	84
D.2.3 Relative Effizienz	85
D.3 Verwendung von Eignungsprüfdaten zur Bewertung der Vergleich- und Wiederholpräzision eines Messverfahrens	86
Anhang E (informativ) Erläuternde Beispiele	87
E.1 Wirkung von zensierten Werten ^N (5.5.3.3)	87
E.2 Homogenitäts- und Stabilitätsprüfung – Arsen (As) in Schokolade (6.1)	88
E.3 Umfassendes Beispiel für Atrazin in Trinkwasser	90
E.4 Umfassendes Beispiel für Quecksilber in Tierfutter	93
E.5 Referenzwert von einem einzelnen Laboratorium: Los-Angeles-Wert von Zuschlagstoffen (7.5)	97
E.6 Beispiel für ein Bootstrap-Verfahren für coliforme Bakterien in Lebensmittelproben (7.7.2)	98
E.7 Vergleich von Referenzwert und Konsens-Mittelwert (7.8)	100
E.8 Bestimmung von Bewertungskriterien durch Erfahrung mit früheren Runden: Toxaphen in Trinkwasser (8.3)	100
E.9 Anhand eines allgemeinen Modells: Horwitz-Gleichung (8.4)	103
E.10 Bestimmung der Leistung aus einem Präzisionsversuch: Bestimmung des Zementgehalts von Festbeton (8.5)	103
E.11 Balkendiagramme von standardisierten systematischen Abweichungskomponenten: Antikörperkonzentrationen (10.4)	104
E.12 Youden-Diagramm — Antikörper-Konzentrationen (10.5)	104
E.13 Diagramm der Wiederholstandardabweichungen: Antikörper-Konzentrationen (10.6)	107
E.14 Grafische Verfahren zur Rückverfolgung der Leistung über die Zeit (10.8)	108
E.15 Qualitative Datenanalyse; Beispiel für eine Ordinalgröße: Hautreaktion auf ein Kosmetikum (Abschnitt 11)	111
Literaturhinweise	113

Nationales Vorwort

Dieses Dokument (ISO 13528:2015) wurde vom Technischen Komitee ISO/TC 69 „Applications of statistical methods“ erarbeitet, dessen Sekretariat von ANSI (USA) gehalten wird.

Das zuständige deutsche Gremium ist Arbeitskreis NA 119-01-03-06-04 AK „Ringversuche zur Eignungsprüfung“ des Arbeitsausschusses NA 119-01-03 AA „Wasseruntersuchung“ im DIN-Normenausschuss Wasserwesen (NAW).

Deutscher Mitträger ist der DIN-Normenausschuss Qualitätsmanagement, Statistik und Zertifizierungsgrundlagen (NQSZ).

Um Zweifelsfälle in der Übersetzung auszuschließen, ist die englische Originalfassung beigelegt. Die Nutzungsbedingungen für den deutschen Text des Norm-Entwurfes gelten gleichermaßen auch für den englischen Text.

Für die in diesem Dokument zitierten internationalen Dokumente wird im Folgenden auf die entsprechenden deutschen Dokumente hingewiesen.

ISO 3534-1	siehe	DIN ISO 3534-1
ISO 3534-2	siehe	DIN ISO 3534-2
ISO 5725-1	siehe	DIN ISO 5725-1
ISO 5725-2	siehe	DIN ISO 5725-2
ISO 5725-3	siehe	DIN ISO 5725-3
ISO 5725-4	siehe	DIN ISO 5725-4
ISO 5725-5	siehe	DIN ISO 5725-5
ISO 5725-6	siehe	DIN ISO 5725-6
ISO 11352	siehe	DIN ISO 11352
ISO 11843-1	siehe	DIN ISO 11843-1
ISO 11843-2	siehe	DIN ISO 11843-2
ISO/IEC 17025	siehe	DIN EN ISO/IEC 17025
ISO/IEC 17043	siehe	DIN EN ISO/IEC 17043

Änderungen

Gegenüber DIN ISO 13528:2009-01 wurden folgende Änderungen vorgenommen:

- Anpassung an ISO/IEC 17043:2010 hinsichtlich der überarbeiteten Struktur, um den Prozess der Planung, Analyse und Berichterstattung von Eignungsprüfungsprogrammen besser zu beschreiben;
- Verzicht auf einige Verfahren, die nicht länger als geeignet angesehen werden;
- neue Abschnitte zu qualitativen Daten und zusätzlichen robusten statistischen Verfahren wurden hinzugefügt;
- die Norm wurde redaktionell überarbeitet.

Nationaler Anhang NA (informativ)

Literaturhinweise

DIN ISO 3534-1, *Statistik — Begriffe und Formelzeichen — Teil 1: Wahrscheinlichkeit und allgemeine statistische Begriffe*

DIN ISO 3534-2, *Statistik — Begriffe und Formelzeichen — Teil 2: Angewandte Statistik*

DIN ISO 5725-1, *Genauigkeit (Richtigkeit und Präzision) von Meßverfahren und Meßergebnissen — Teil 1: Allgemeine Grundlagen und Begriffe*

DIN ISO 5725-2, *Genauigkeit (Richtigkeit und Präzision) von Messverfahren und Messergebnissen — Teil 2: Grundlegende Methode für Ermittlung der Wiederhol- und Vergleichpräzision eines vereinheitlichten Messverfahrens*

DIN ISO 5725-3, *Genauigkeit (Richtigkeit und Präzision) von Messverfahren und Messergebnissen — Teil 3: Präzisionsmaße eines vereinheitlichten Messverfahrens unter Zwischenbedingungen*

DIN ISO 5725-4, *Genauigkeit (Richtigkeit und Präzision) von Messverfahren und Messergebnissen — Teil 4: Grundlegende Methoden für die Ermittlung der Richtigkeit eines vereinheitlichten Messverfahrens*

DIN ISO 5725-5, *Genauigkeit (Richtigkeit und Präzision) von Messverfahren und Messergebnissen — Teil 5: Alternative Methoden für die Ermittlung der Präzision eines vereinheitlichten Messverfahrens*

DIN ISO 5725-6, *Genauigkeit (Richtigkeit und Präzision) von Messverfahren und Messergebnissen — Teil 6: Anwendung von Genauigkeitswerten in der Praxis*

DIN ISO 11352, *Wasserbeschaffenheit — Abschätzung der Messunsicherheit beruhend auf Validierungs- und Kontrolldaten*

DIN ISO 11843-1, *Erkennungsfähigkeit — Teil 1: Begriffe*

DIN ISO 11843-2, *Erkennungsfähigkeit — Teil 2: Verfahren im Fall der linearen Kalibrierung*

DIN EN ISO/IEC 17025, *Allgemeine Anforderungen an die Kompetenz von Prüf- und Kalibrierlaboratorien*

DIN EN ISO/IEC 17043, *Konformitätsbewertung — Allgemeine Anforderungen an Eignungsprüfungen*

Vorwort

ISO (die Internationale Organisation für Normung) ist eine weltweite Vereinigung von Nationalen Normungsorganisationen (ISO-Mitgliedsorganisationen). Die Erstellung von Internationalen Normen wird normalerweise von ISO Technischen Komitees durchgeführt. Jede Mitgliedsorganisation, die Interesse an einem Thema hat, für welches ein Technisches Komitee gegründet wurde, hat das Recht, in diesem Komitee vertreten zu sein. Internationale Organisationen, staatlich und nicht-staatlich, in Liaison mit ISO, nehmen ebenfalls an der Arbeit teil. ISO arbeitet eng mit der Internationalen Elektrotechnischen Kommission (IEC) bei allen elektrotechnischen Themen zusammen.

Die Verfahren, die bei der Entwicklung dieses Dokuments angewendet wurden und die für die weitere Pflege vorgesehen sind, werden in den ISO/IEC-Direktiven, Teil 1 beschrieben. Im Besonderen sollten die für die verschiedenen ISO-Dokumentenarten notwendigen Annahmekriterien beachtet werden. Dieses Dokument wurde in Übereinstimmung mit den Gestaltungsregeln der ISO/IEC-Direktiven, Teil 2 erarbeitet (siehe www.iso.org/directives).

Es wird auf die Möglichkeit hingewiesen, dass einige Elemente dieses Dokuments Patentrechte berühren können. ISO ist nicht dafür verantwortlich, einige oder alle diesbezüglichen Patentrechte zu identifizieren. Details zu allen während der Entwicklung des Dokuments identifizierten Patentrechten finden sich in der Einleitung und/oder in der ISO-Liste der empfangenen Patenterklärungen (siehe www.iso.org/patents).

Jeder in diesem Dokument verwendete Handelsname wird als Information zum Nutzen der Anwender angegeben und stellt keine Anerkennung dar.

Eine Erläuterung der Bedeutung ISO-spezifischer Benennungen und Ausdrücke, die sich auf Konformitätsbewertung beziehen, sowie Informationen über die Beachtung der Grundsätze der Welthandelsorganisation (WTO) zu technischen Handelshemmnissen (TBT, en: Technical Barriers to Trade) durch ISO enthält der folgende Link: www.iso.org/iso/foreword.html.

Das für dieses Dokument verantwortliche Komitee ist ISO/TC 69, *Applications of statistical methods*, Unterkomitee SC 6, *Measurement methods and results*.

Diese zweite Ausgabe von ISO 13528 ersetzt die erste Ausgabe (ISO 13528:2005), die technisch überarbeitet wurde. Diese zweite Ausgabe enthält Änderungen, um das Dokument in Einklang mit ISO/IEC 17043:2010 zu bringen, die ISO Guide 43-1:1997 ersetzt. Es folgt einer überarbeiteten Struktur, um den Prozess der Planung, Analyse und Berichterstattung von Eignungsprüfungsprogrammen besser zu beschreiben. Es wurden auch einige Verfahren entfernt, die nicht länger als geeignet angesehen werden und einige neue Abschnitte hinzugefügt oder bearbeitet, um sie in Einklang mit ISO/IEC 17043 zu bringen, sowie um Klarheit zu schaffen und zur Korrektur geringfügiger Fehler. Neue Abschnitte zu qualitativen Daten und zusätzlichen robusten statistischen Verfahren wurden hinzugefügt.

Diese korrigierte Fassung von ISO 13528:2015 enthält folgenden Korrekturen:

- 7.5.2.2, Gleichung (5): „ U_{char} “ wurde ersetzt durch „ uc_{char} “;
- 9.8.2, ANMERKUNG: Verweisung auf „E.3“ wurde ersetzt durch eine Verweisung auf „E.4“;
- 10.3.2, Listenpunkt iv), Gleichung (19): ein Additionszeichen wurde zwischen „ q_{min} “ und „ $(i-1)$ “ hinzugefügt;
- B.2.3, b), Zeile vor Tabelle B.1: die Verweisung „[33]“ wurde durch die Verweisung „[32]“ ersetzt;

- B.2.3, Tabelle B.1, erste Zeile, erste Spalte: „ gm “ wurde ersetzt durch „ g “;
- B.2.3, Absatz unter Tabelle B.1: in der ersten Gleichung wurde „ F_{2m} “ ersetzt durch „ F_m “ und die Tiefstellung von „ -1 “ wurde entfernt; in der zweiten Gleichung wurde die Tiefstellung von „ -1 “ entfernt;
- B.2.3, b), ANMERKUNG: die Gleichung für „ F_1 “ wird nun durch „ $(g-1)$ “ dividiert;
- B.3, Gleichungen (B.7), (B.14) und (B.16): die Terme s_x^2 und s_x wurden ersetzt durch s_x^2 und $s_{\bar{x}}$; zusätzlich wurde in Gleichung (B.16) das Quadratwurzelzeichen aus der Klammer „ $\max_x(0, \dots)$ “ herausgenommen und ihr außen vorangestellt;
- B.3, Gleichung (B.8): „ s_t^2 “ wurde ersetzt durch „ w_t^2 “;
- C.3.1, ANMERKUNG 2, erste Zeile: der Artikel wurde vor „identisch“ entfernt;
- C.3.1, Absatz nach Gleichung (C.10), zweite Zeile: die Worte „die modifizierten Daten in“ wurden gelöscht;
- C.3.1, letzte Anmerkung: „ANMERKUNG“ wurde ersetzt durch „ANMERKUNG 3“ und die Verweisung auf „E.3 und E.4“ wurde durch eine Verweisung auf „E.1 und E.3“ ersetzt;

Die folgenden geringfügigen redaktionellen Berichtigungen wurden aus Gründen der Konsistenz im gesamten Dokument umgesetzt:

- 8.3.1, dritter Spiegelstrich, letzte Zeile: das erste Auftreten von „zugelassene“ (nach „mehr“) wurde gelöscht;
- 8.6.1, erste Zeile: „ σ_{pt} “ wurde ersetzt durch „ (σ_{pt}) “ (für eine einheitliche Darstellung);
- B.4.1.2, zweiter Spiegelstrich, zweite Zeile: das Wort „Proben“ wurde ersetzt durch „Prüfgegenstände für die Eignungsprüfung“ (für terminologische Konsistenz);
- Anhänge D und E, Titel: die Anfangsbuchstaben aller Worte nach dem ersten sind nun in Kleinbuchstaben gesetzt (für eine einheitliche Darstellung).

0 Einleitung

0.1 Zweck von Eignungsprüfungen

Eignungsprüfungen umfassen die Anwendung von Ringversuchen, um die Leistung von Teilnehmern (das können Laboratorien, Prüfstellen oder Einzelpersonen sein) bei bestimmten Prüfungen oder Messungen zu bestimmen und ihre andauernde Leistung zu überwachen. Es gibt eine Reihe von üblichen Zwecken der Eignungsprüfung, wie in ISO/IEC 17043:2010, Einleitung, beschrieben. Diese schließen die Bewertung der Leistung von Laboratorien, die Identifizierung von Problemen in Laboratorien, Festlegungen zu Wirksamkeit und Vergleichbarkeit von Prüf- oder Messverfahren, die Schaffung zusätzlichen Vertrauens für Kunden von Laboratorien, die Validierung von Unsicherheitsangaben und die Ausbildung von teilnehmenden Laboratorien ein. Die angewendete statistische Planung und die Analysetechniken müssen für den angegebenen Zweck geeignet sein.

0.2 Begründung für die Bewertung durch Scores in Eignungsprüfungsprogrammen

Für die Bewertung durch Scores steht eine Vielzahl von Strategien zur Verfügung, die bei Eignungsprüfungen angewendet werden. Obwohl sich die Detailberechnungen unterscheiden, vergleichen die meisten Eignungsprüfungsprogramme die Abweichung des Teilnehmers von einem zugewiesenen Wert mit einem numerischen Kriterium, das für die Entscheidung verwendet wird, ob die Abweichung einen Grund zur Besorgnis darstellt oder nicht. Die Strategien, die bei der Wertzuweisung und bei der Auswahl eines Kriteriums zur Beurteilung der Teilnehmerabweichungen angewendet werden, sind deshalb entscheidend. Besonders wichtig ist, zu berücksichtigen, ob der zugewiesene Wert und das Kriterium zur Beurteilung von Abweichungen unabhängig von Teilnehmerergebnissen sein sollten oder aus den übermittelten Ergebnissen abgeleitet werden sollten. In dieser Norm werden beide Strategien berücksichtigt. Allerdings wird auf die Diskussion der Vor- und Nachteile der Wahl von zugewiesenen Werten oder von Kriterien für die Beurteilung von Abweichungen, die nicht aus Teilnehmerergebnissen stammen, hingewiesen, die den Abschnitten 7 und 8 entnommen werden kann. Es zeigt sich, dass die Auswahl zugewiesener Werte und Beurteilungskriterien unabhängig von den Teilnehmerergebnissen im Allgemeinen Vorteile bietet. Das trifft besonders auf das Kriterium zu, das für die Beurteilung von Abweichungen vom zugewiesenen Wert verwendet wird – wie die Standardabweichung für die Eignungsbeurteilung oder eine Vorgabe für die Messabweichung – für die eine einheitliche Wahl auf Grundlage der Eignung für einen bestimmten Verwendungszweck der Messergebnisse besonders zweckmäßig ist.

0.3 ISO 13528 und ISO/IEC 17043

ISO 13528 bietet Unterstützung bei der Umsetzung von ISO/IEC 17043, insbesondere bei den Anforderungen an die statistische Planung, Validierung von Prüfgegenständen für die Eignungsprüfung, Überprüfung der Ergebnisse und die Berichterstattung von zusammenfassenden Kenngrößen. ISO/IEC 17043:2010, Anhang B, beschreibt kurz die allgemeinen statistischen Verfahren, die in Eignungsprüfungsprogrammen Anwendung finden. Diese Internationale Norm soll ISO/IEC 17043 ergänzen, indem sie einen detaillierten Leitfaden bereitstellt, der in jenem Dokument bezüglich besonderer statistischer Verfahren für die Eignungsprüfung fehlt.

Die Definition von Eignungsprüfung nach ISO/IEC 17043 wird in ISO 13528 wiederholt, mit Anmerkungen, die verschiedene Arten von Eignungsprüfungen beschreiben sowie die Bandbreite an Planungen, die verwendet werden können. Diese Norm kann nicht sämtliche Zwecke, Planungen, Matrices und Messgrößen spezifisch behandeln. Die Verfahren nach ISO 13528 sollen allgemein anwendbar sein, besonders auf neu erstellte Eignungsprüfungsprogramme. Es wird erwartet, dass statistische Verfahren, die bei einem bestimmten Eignungsprüfungsprogramm angewendet werden, sich mit fortschreitendem Reifegrad des Programms weiterentwickeln; und dass die Scores, Bewertungskriterien und grafischen Verfahren verfeinert werden, um den speziellen Erfordernissen einer Zielgruppe von Teilnehmern, Akkreditierungsstellen und regelsetzenden Behörden besser gerecht zu werden.

ISO 13528 schließt bereits veröffentlichte Leitfäden für die Eignungsprüfung von chemischen Analyselaboratorien ein [32], aber sie enthält darüber hinaus eine breitere Auswahl an Verfahren, um die Anwendung mit gültigen Messverfahren und qualitativer Erkennung zu ermöglichen. Diese Überarbeitung von ISO 13528:2005 enthält die meisten der statistischen Verfahren und Anleitungen aus der ersten Ausgabe, die soweit notwendig um die vorstehend verwiesenen Dokumente und den erweiterten Anwendungsbereich von ISO/IEC 17043 erweitert wurde. ISO/IEC 17043 umfasst Eignungsprüfungen für Einzelpersonen und Inspektionsstellen sowie Anhang B, der Überlegungen zu qualitativen Ergebnissen enthält.

Diese Norm enthält statistische Verfahren, die in Einklang mit anderen Internationalen Normen stehen, speziell denen von TC 69, SC 6, insbesondere der Normenreihe ISO 5725, *Accuracy: trueness and precision*. Die Verfahren sollen auch, soweit anwendbar, andere Internationale Normen widerspiegeln, und sie sollen in Einklang stehen mit ISO/IEC Guide 98-3 (GUM) und ISO/IEC Guide 99 (VIM).

0.4 Statistischer Sachverstand

Nach ISO/IEC 17043:2010 wird gefordert, dass ein Anbieter von Eignungsprüfungen um kompetent zu sein, Zugang zu statistischem Fachwissen besitzen muss, und dass er bestimmtes Personal zur Durchführung statistischer Analysen autorisieren muss. Weder ISO/IEC 17043 noch diese Internationale Norm können darüber hinausgehend festlegen, was dieses erforderliche Fachwissen ist. Bei einigen Anwendungen ist ein höherer akademischer Grad auf dem Gebiet der Statistik zweckmäßig, aber üblicherweise können die Erfordernisse an den Sachverstand auch von Einzelpersonen mit technischem Fachwissen in anderen Bereichen erfüllt werden, die mit den grundlegenden statistischen Konzepten vertraut sind und die in den üblichen Verfahren, die auf die Analyse von Daten aus Eignungsprüfungsprogrammen anwendbar sind, erfahren oder geschult sind. Falls eine Einzelperson mit der statistischen Planung und/oder Analyse betraut wird, ist es sehr wichtig, dass diese Person in Ringversuchen erfahren ist, sogar wenn diese Person einen höheren akademischen Grad in Statistik besitzt. Konventionelle Schulung für höhere Statistik umfasst häufig keine Übungen mit Ringversuchen, und die speziellen Fälle von Messabweichungen, die bei Eignungsprüfungen auftreten, können undurchsichtig erscheinen. Der Leitfaden in dieser Internationalen Norm kann nicht das gesamte erforderliche Fachwissen bereitstellen, um sämtliche Anwendungen zu berücksichtigen, und er kann nicht die Erfahrung ersetzen, die bei der Arbeit mit Ringversuchen gesammelt wurde.

0.5 Computersoftware

Computersoftware, die bei der statistischen Analyse von Eignungsprüfungsdaten benötigt wird, kann stark variieren, angefangen bei einfachen arithmetischen Tabellenkalkulationen für kleine Eignungsprüfungsprogramme, die bekannte Referenzwerte verwenden bis hin zu hochentwickelter Statistik-Software, die für statistische Verfahren eingesetzt wird, die auf iterative Berechnungen oder andere fortgeschrittene numerische Verfahren angewiesen sind. Die meisten der Verfahren nach dieser Internationalen Norm können mit konventionellen Tabellenkalkulationsanwendungen ausgeführt werden, möglicherweise mit individuell angepassten Routinen für ein bestimmtes Programm oder eine bestimmte Analyse; einige Verfahren werden rechnergestützte Anwendungen erfordern, die frei verfügbar sind (zum Zeitpunkt der Veröffentlichung dieser Norm). In jedem Fall sollten Anwender die Genauigkeit ihrer Berechnungen verifizieren, besonders wenn spezielle Routinen vom Anwender eingeführt wurden. Sogar wenn die Verfahren nach dieser Internationalen Norm geeignet sind und mit entsprechenden rechnergestützten Anwendungen korrekt umgesetzt werden, können sie nicht ohne Beaufsichtigung durch eine Person mit technischem und statistischem Sachverstand angewendet werden, der ausreicht, um Anomalien zu erkennen und zu untersuchen, die in jeder Eignungsprüfungsrunde auftreten können.

1 Anwendungsbereich

Diese Internationale Norm enthält detaillierte Beschreibungen von statistischen Verfahren für Anbieter von Eignungsprüfungen, damit sie diese zur Planung von Eignungsprüfungsprogrammen und zur Auswertung der aus diesen Programmen erhaltenen Daten anwenden. Diese Norm enthält Empfehlungen zur Interpretation von Daten aus Eignungsprüfungen durch Teilnehmer an derartigen Programmen und durch Akkreditierungsstellen.

Die Verfahren in dieser Norm können angewendet werden, um nachzuweisen, dass die von Laboratorien, Prüfstellen und Einzelpersonen erzielten Messergebnisse den festgelegten Kriterien für akzeptable Leistung entsprechen.

Diese Norm gilt für Eignungsprüfungen, bei denen die berichteten Ergebnisse entweder quantitative Messungen oder qualitative Beobachtungen von Prüfgegenständen sind.

ANMERKUNG Die Verfahren nach dieser Norm können auch auf die Beurteilung von Expertenmeinungen anwendbar sein, sofern die Meinungen oder Einschätzungen in einer Form berichtet werden, die mit einem unabhängigen Referenzwert oder einer Konsenskenngroße objektiv vergleichbar ist. Zum Beispiel können die Festlegungen dieser Norm zu nominalen (qualitativen) Eigenschaften gelten, wenn Prüfgegenstände für die Eignungsprüfung nach bekannten Kategorien klassifiziert werden – oder bei der Bestimmung durch Prüfung, ob Prüfgegenstände für die Eignungsprüfung derselben Ursprungsquelle entstammen oder nicht – und wenn die Klassifizierungsergebnisse objektiv verglichen werden.

2 Normative Verweisungen

Die folgenden Dokumente, die in diesem Dokument teilweise oder als Ganzes zitiert werden, sind für die Anwendung dieses Dokuments erforderlich. Bei datierten Verweisungen gilt nur die in Bezug genommene Ausgabe. Bei undatierten Verweisungen gilt die letzte Ausgabe des in Bezug genommenen Dokuments (einschließlich aller Änderungen).

ISO Guide 30, *Reference materials — Selected terms and definitions*

ISO 3534-1, *Statistics — Vocabulary and symbols — Part 1: General statistical terms and terms used in probability*

ISO 3534-2, *Statistics — Vocabulary and symbols — Part 2: Applied statistics*

ISO 5725-1, *Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions*

ISO/IEC 17043, *Conformity assessment — General requirements for proficiency testing*

ISO/IEC Guide 99, *International vocabulary of metrology — Basic and general concepts and associated terms (VIM)*

3 Begriffe

Für die Anwendung dieses Dokuments gelten die Begriffe nach ISO 3534-1, ISO 3534-2, ISO 5725-1, ISO/IEC 17043, ISO/IEC Guide 99, ISO Guide 30, und die folgenden Begriffe. Im Fall von Abweichungen zwischen diesen normativen Verweisungen bei der Verwendung von Begriffen, gelten die Begriffe nach ISO 3534, Teile 1 und 2. Mathematische Zeichen sind in Anhang A aufgeführt.

3.1

Ringversuch

Organisation, Durchführung und Bewertung von Messungen oder Prüfungen an denselben oder ähnlichen Gegenständen durch zwei oder mehr Laboratorien entsprechend vorherbestimmten Bedingungen

3.2

Eignungsprüfung

Bewertung der Leistung eines Teilnehmers durch Ringversuche nach zuvor aufgestellten Kriterien

Anmerkung 1 zum Begriff: Für die Anwendung dieser Internationalen Norm wird der Begriff „Eignungsprüfung“ im seinem weitesten Sinne verwendet und umfasst unter anderem:

- quantitatives Programm — Ziel ist es, eine oder mehrere Messgrößen für jeden Prüfgegenstand für die Eignungsprüfung zu quantifizieren;
- qualitatives Programm — Ziel ist es, ein oder mehrere qualitative Merkmale des Prüfgegenstands zu identifizieren oder zu beschreiben;
- sequentielles Programm — wenn ein oder mehrere Prüfgegenstände für Prüfungen oder Messungen der Reihe nach verteilt und in zeitlichen Abständen zum Anbieter für Eignungsprüfungen zurückgesendet werden;
- zeitgleiches Programm — wenn Prüfgegenstände für gleichzeitig verlaufende Prüfungen oder Messungen innerhalb einer festgelegten Zeitspanne verteilt werden;
- einmalige Prüfung — wenn Prüfgegenstände einmalig zur Verfügung gestellt werden;
- fortlaufendes Schema — wobei Prüfgegenstände für die Eignungsprüfung in regelmäßigen Zeitabständen zur Verfügung gestellt werden;
- Probenahme — wenn Proben für nachfolgende Analysen genommen werden und der Zweck des Eignungsprüfungsprogramms die Bewertung der Ausführung der Probenahme umfasst; und
- Dateninterpretation — wenn Datenreihen oder Informationsreihen bereitgestellt werden und die Informationen aufbereitet werden, um eine Interpretation zu liefern (oder andere Ergebnisse).

3.3

zugewiesener Wert

Wert, der einem bestimmten Merkmal eines Prüfgegenstandes für die Eignungsprüfung zugewiesen wurde

3.4

Standardabweichung für die Eignungsbeurteilung

Maß der Streuung, das bei der Bewertung der Ergebnisse aus der Eignungsprüfung verwendet wird

Anmerkung 1 zum Begriff: Das kann als Standardabweichung der Grundgesamtheit der Ergebnisse von einer hypothetischen Grundgesamtheit von Laboratorien, die in exakter Übereinstimmung mit den Anforderungen arbeiten, interpretiert werden.

Anmerkung 2 zum Begriff: Die Standardabweichung für die Eignungsbeurteilung bezieht sich nur auf Ergebnisse auf der Verhältnisskala und auf der Intervallskala.

Anmerkung 3 zum Begriff: Nicht alle Eignungsprüfungsprogramme bewerten die Leistung auf der Grundlage der Streuung der Ergebnisse.

[QUELLE: ISO/IEC 17043:2010, modifiziert — In der Definition wurde „basierend auf verfügbaren Informationen“ gelöscht. Anmerkung 1 zum Begriff: wurde hinzugefügt, Anmerkungen 2 und 3 zum Begriff wurden leicht überarbeitet.]

3.5

Messabweichung

Messwert minus eines Referenzwerts

[QUELLE: ISO/IEC Guide 99:2007, modifiziert — Anmerkungen wurden gelöscht.]

3.6

Grenzwert der Messabweichung

Extremwert einer Messabweichung in Bezug auf einen bekannten Referenzwert, durch Spezifikationen oder Vorschriften zugelassen für eine Messung, ein Messgerät oder ein Messsystem

[QUELLE: ISO/IEC Guide 99:2007, modifiziert — Anmerkungen wurden gelöscht.]

3.7

z-Score

standardisiertes Maß der Leistung, berechnet unter Verwendung des Teilnehmerergebnisses, des zugewiesenen Werts und der Standardabweichung für die Eignungsbeurteilung

Anmerkung 1 zum Begriff: Eine häufige Variante des z-Scores, mitunter als z' bezeichnet (üblicherweise als z-Strich gesprochen) wird durch die Kombination der Unsicherheit des zugewiesenen Werts mit der Standardabweichung für die Eignungsbeurteilung vor Berechnung des z-Scores gebildet.

3.8

zeta-Score

standardisiertes Maß für die Leistung, berechnet unter Verwendung des Teilnehmerergebnisses, des zugewiesenen Werts und der kombinierten Standardunsicherheiten für das Ergebnis und den zugewiesenen Wert

3.9

Anteil des zulässigen Grenzscores

standardisiertes Maß für die Leistung, berechnet unter Verwendung des Teilnehmerergebnisses, des zugewiesenen Werts und des Kriteriums für die Messabweichung in einer Eignungsprüfung

Anmerkung 1 zum Begriff: Bei Einzelergebnissen kann die Leistung als die Abweichung vom zugewiesenen Wert (D oder $D\%$) angegeben werden.

3.10

Eingriffssignal

Hinweis auf Handlungsbedarf infolge eines Eignungsprüfungsergebnisses

BEISPIEL Ein z-Score, der 2 überschreitet, wird üblicherweise als Anzeichen gewertet, dass Bedarf zur Untersuchung möglicher Ursachen besteht; ein z-Score, der 3 überschreitet, wird üblicherweise als ein Eingriffssignal gewertet, das Bedarf für eine Korrekturmaßnahme anzeigt.

3.11

Konsenswert

Wert, der aus einer Sammlung von Ergebnissen in einem Ringversuch abgeleitet wurde

Anmerkung 1 zum Begriff: Der Begriff „Konsenswert“ wird üblicherweise verwendet, um Schätzwerte der Lage und der Streuung zu beschreiben, die aus Teilnehmerergebnissen in einer Eignungsprüfungsrunde abgeleitet sind; er kann jedoch auch auf Werte bezogen werden, die aus Ergebnissen einer festgelegten Teilmenge derartiger Ergebnisse abgeleitet sind, z. B. aus einer Anzahl von Expertenlaboratorien.

3.12

Ausreißer

Wert unter einer Serie von Werten, der mit den anderen Werten dieser Serie nicht verträglich ist

Anmerkung 1 zum Begriff: Ein Ausreißer kann zufällig aus der erwarteten Grundgesamtheit entstehen, aus einer unterschiedlichen Grundgesamtheit stammen oder das Ergebnis einer fehlerhaften Aufzeichnung oder anderer grober Fehler sein.

Anmerkung 2 zum Begriff: Viele Eignungsprüfungsprogramme verwenden den Begriff „Ausreißer“, um ein Ergebnis zu bezeichnen, dass ein Eingriffssignal erzeugt. Das ist nicht der bestimmungsgemäße Gebrauch des Begriffs. Während Ausreißer üblicherweise Eingriffssignale erzeugen werden, ist es möglich, dass Eingriffssignale von Ergebnissen herrühren, die keine Ausreißer sind.

[QUELLE: ISO 5725-1:1994, modifiziert – Die Anmerkungen zum Begriff wurden hinzugefügt.]

3.13

Teilnehmer

Laboratorium, Organisation oder Einzelperson, welche(s) die Prüfgegenstände erhält und Ergebnisse für die Bewertung durch den Anbieter der Eignungsprüfung abliefern

3.14

Prüfgegenstand für die Eignungsprüfung

Probe, Produkt, Artefakt, Referenzmaterial, Teil einer Einrichtung, Normal, Datenreihe oder andere Informationen, die für die Beurteilung der Leistung von Teilnehmern in einer Eignungsprüfung verwendet werden

Anmerkung 1 zum Begriff: In den meisten Fällen entsprechen Prüfgegenstände für die Eignungsprüfung der Definition von „Referenzmaterial“ nach ISO Guide 30 (3.17).

3.15

Anbieter von Eignungsprüfungen

Organisation, die die Verantwortung für alle Aufgaben bezüglich Entwicklung und Durchführung eines Eignungsprüfungsprogramms übernimmt

3.16

Eignungsprüfungsprogramm

Eignungsprüfung, entwickelt und durchgeführt in einer oder mehreren Runden für ein spezifisches Prüf-, Mess-, Kalibrier- oder Inspektionsgebiet

Anmerkung 1 zum Begriff: Ein Eignungsprüfungsprogramm kann eine bestimmte Art von Prüfung, Kalibrierung, Inspektion oder eine Anzahl von Prüfungen, Kalibrierungen oder Inspektionen zu Prüfgegenständen zur Eignungsprüfung beinhalten.

3.17

Referenzmaterial

RM

Material, das im Hinblick auf eine oder mehrere festgelegte Eigenschaft(en) ausreichend homogen und stabil ist und hergestellt wurde, um für dessen vorgesehene Verwendung in einem Messverfahren geeignet zu sein

Anmerkung 1 zum Begriff: RM ist ein allgemeiner Begriff.

Anmerkung 2 zum Begriff: Eigenschaften können quantitativ oder qualitativ sein, z. B. Identität von Substanzen oder Arten.

Anmerkung 3 zum Begriff: Verwendungen können die Kalibrierung eines Messsystems, die Beurteilung eines Messverfahrens, das Zuweisen von Werten zu anderen Materialien und die Qualitätsregelung einschließen.

[QUELLE: ISO Guide 30:2015, modifiziert – Anmerkung 4 wurde gelöscht.]

3.18

zertifiziertes Referenzmaterial

ZRM

Referenzmaterial, das durch ein metrologisch valides Verfahren in Bezug auf eine oder mehrere festgelegte Eigenschaft(en) charakterisiert und durch ein RM Zertifikat begleitet wird, in dem der Wert der festgelegten Eigenschaft, die zugehörige Unsicherheit und eine Erklärung der messtechnischen Rückführung angegeben werden

Anmerkung 1 zum Begriff: Der Begriff „Wert“ schließt ein Nominalmerkmal oder qualitative Attribute ein, wie Identität oder Reihenfolge. Unsicherheiten für derartige Attribute können als Wahrscheinlichkeiten oder Vertrauensniveaus angegeben werden.

[QUELLE: ISO Guide 30:2015, modifiziert – Anmerkungen 2, 3 und 4 wurden gelöscht.]

4 Allgemeine Grundsätze

4.1 Allgemeine Anforderungen an statistische Verfahren

4.1.1 Die angewendeten statistischen Verfahren müssen gebrauchstauglich und statistisch gültig sein. Alle statistischen Annahmen auf denen die Verfahren oder die Planung beruhen, müssen in der Planung oder in einer schriftlichen Beschreibung des Eignungsprüfungsprogramms angegeben werden, und es muss nachgewiesen werden, dass diese Annahmen begründet sind.

ANMERKUNG Ein statistisch gültiges Verfahren besitzt eine fundierte theoretische Grundlage, eine bekannte Leistung unter erwarteten Anwendungsbedingungen und beruht auf Annahmen oder Bedingungen, für die nachgewiesen werden kann, dass sie in ausreichendem Maß für die jeweiligen Zwecke auf die Daten zutreffen.

4.1.2 Die statistische Planung und die Datenanalyseverfahren müssen mit den festgelegten Zielen für das Eignungsprüfungsprogramm übereinstimmen.

4.1.3 Der Anbieter von Eignungsprüfungen muss Teilnehmern eine Beschreibung der angewendeten Berechnungsverfahren, eine Erklärung der allgemeinen Interpretation der Ergebnisse und eine Aussage zu allen Beschränkungen in Bezug auf die Interpretation bereitstellen. Diese muss entweder in jedem Bericht für jede Runde des Eignungsprüfungsprogramms verfügbar sein oder in einer gesonderten Zusammenfassung der Verfahren, das den Teilnehmern zur Verfügung steht.

4.1.4 Der Anbieter von Eignungsprüfungen muss sicherstellen, dass die gesamte Software angemessen validiert wird.

4.2 Grundmodell

4.2.1 Für quantitative Ergebnisse in Eignungsprüfungsprogrammen, bei denen ein Einzelergebnis für einen bestimmten Prüfgegenstand berichtet wird, wird das Grundmodell in Gleichung (1) angegeben.

$$x_i = \mu + \varepsilon_i \quad (1)$$

Dabei ist

x_i = Ergebnis der Eignungsprüfung von Teilnehmer i ;

μ = wahrer Wert einer Messgröße;

ε_i = Messabweichung für den Teilnehmer i , mit einer Verteilung nach einem relevanten Modell.

ANMERKUNG 1 Übliche Modelle für ε umfassen: die Normalverteilung $\varepsilon_i \sim N(0, \sigma^2)$ mit dem Mittelwert 0 und einer Varianz, die entweder konstant ist oder für jedes Laboratorium verschieden; oder häufiger, eine mit „Ausreißern kontaminierte Normalverteilung“, bestehend aus einer Mischung einer Normalverteilung mit einer breiteren Verteilung, die die Grundgesamtheit abweichender Ergebnisse repräsentiert.

ANMERKUNG 2 Die Grundlage der Leistungsbewertung mit z-Scores und σ_{pt} besteht darin, dass in einer „idealisierten“ Grundgesamtheit kompetenter Laboratorien die Standardabweichung zwischen den Laboratorien σ_{pt} oder weniger betragen würde.

ANMERKUNG 3 Dieses Modell unterscheidet sich vom Grundmodell nach ISO 5725 darin, dass es nicht die systematische Abweichungskomponente eines Labors, B_p , einschließt. Das liegt darin begründet, dass die systematische Abweichungskomponente eines Labors nicht von übrigen Abweichungen unterschieden werden kann, wenn nur eine einzelne Beobachtung berichtet wird. Wenn die Ergebnisse eines Teilnehmers aus mehreren Runden oder von mehreren Prüfgegenständen berücksichtigt werden, kann es jedoch nützlich sein, einen gesonderten Term für die systematische Abweichungskomponente eines Labors einzubeziehen.

4.2.2 Für Ordinalergebnisse oder qualitative Ergebnisse können andere Modelle geeignet sein oder es könnte kein statistisches Modell existieren.

4.3 Allgemeine Ansätze für die Leistungsbewertung

4.3.1 Es gibt drei verschiedene allgemeine Ansätze für die Leistungsbewertung in einem Eignungsprüfungsprogramm. Mit diesen Ansätzen werden verschiedene Zwecke im Eignungsprüfungsprogramm erfüllt. Die Ansätze sind nachstehend aufgeführt:

- a) Bewertung der Leistung durch Vergleich mit extern erhaltenen Kriterien;
- b) Bewertung der Leistung durch Vergleich mit anderen Teilnehmern;
- c) Bewertung der Leistung durch Vergleich mit der angegebenen Messunsicherheit.

4.3.2 Die allgemeinen Ansätze können in unterschiedlicher Weise zur Bestimmung des zugewiesenen Werts und zur Bestimmung der Kriterien für die Leistungsbewertung verwendet werden; z. B. wenn der zugewiesene Wert der robuste Mittelwert der Teilnehmerergebnisse ist und die Leistungsbewertung aus σ_{pt} oder δ_E abgeleitet wird, wobei δ_E eine vorher festgelegte zulässige Messabweichung ist und $\sigma_{pt} = \delta_E / 3$; gleichermaßen kann in einigen Fällen der zugewiesene Wert ein Referenzwert sein, aber σ_{pt} kann eine robuste Standardabweichung von Teilnehmerergebnissen sein. Bei Ansatz c), der die Messunsicherheit verwendet, ist der zugewiesene Wert üblicherweise ein geeigneter Referenzwert.

5 Leitlinien für die statistische Planung von Eignungsprüfungsprogrammen

5.1 Einleitung zur statistischen Planung von Eignungsprüfungsprogrammen

Eignungsprüfungen befassen sich mit der Beurteilung der Leistung der Teilnehmer und behandeln deshalb nicht speziell die systematische Abweichungskomponente oder Präzision (obwohl diese mit speziellen Planungen beurteilt werden können). Die Leistung der Teilnehmer wird durch statistische Bewertung ihrer Ergebnisse nach den Messungen oder Interpretationen beurteilt, die sie an Prüfgegenständen für die Eignungsprüfung vornehmen. Leistung wird häufig in Form von Leistungswerten ausgedrückt, die eine konsistente Interpretation über eine Spannweite von Messgrößen ermöglichen und den Vergleich von Ergebnissen, die für verschiedene Messgrößen erzielt wurden, auf gleicher Grundlage ermöglichen können. Leistungswerte werden üblicherweise durch Vergleichen der Differenz zwischen einem berichteten Teilnehmerergebnis und einem zugewiesenen Wert mit zulässiger Abweichung oder mit einem Schätzwert der Messunsicherheit der Differenz erhalten. Die Überprüfung der Leistungswerte über mehrere Runden eines Eignungsprüfungsprogramms kann Informationen dazu liefern, ob einzelne Laboratorien Hinweise auf konsistente systematische Einflüsse („systematische Abweichungskomponente“) oder schlechte Langzeitpräzision zeigen.

Die folgenden Abschnitte 5 bis 10 enthalten einen Leitfaden für die Planung quantitativer Eignungsprüfungsprogramme und zur statistischen Behandlung der Ergebnisse, einschließlich Berechnung und Interpretation der verschiedenen Leistungswerte. Abschnitt 11 enthält Überlegungen zu qualitativen Eignungsprüfungsprogrammen (einschließlich Ordinalprogrammen).

5.2 Grundlage einer statistischen Planung

5.2.1 Nach ISO/IEC 17043, 4.4.4.1, müssen statistische Planungen „so entwickelt werden, dass sie den Programmzielen, basierend auf der Art der Daten (quantitative oder qualitative, einschließlich ordinale und nominale), auf statistischen Annahmen, der Art der Abweichung und der erwarteten Anzahl an Ergebnissen gerecht werden“. Daher könnten Eignungsprüfungsprogramme mit unterschiedlichen Zielen und mit verschiedenen Abweichungsquellen, unterschiedlich geplant sein.

Überlegungen zur Versuchsplanung für übliche Ziele sind nachstehend aufgeführt. Andere Ziele sind möglich.

BEISPIEL 1: Die Planung eines Eignungsprüfungsprogramms erfordert für den Vergleich eines Teilnehmerergebnisses mit einem vorbestimmten Referenzwert, innerhalb von Grenzen, die vor Rundenbeginn festgelegt werden, ein Verfahren für das Erzielen eines extern festgelegten Referenzwerts, ein Verfahren zum Festlegen von Grenzwerten und ein Verfahren zur Bewertung durch Scores;

BEISPIEL 2 Die Planung eines Eignungsprüfungsprogramms erfordert für den Vergleich eines Teilnehmerergebnisses mit den kombinierten Ergebnissen einer Gruppe in derselben Runde und in den Grenzen die vor Rundenbeginn festgelegt werden, dass berücksichtigt wird, wie der zugewiesene Wert aus den kombinierten Ergebnissen bestimmt werden wird, sowie welche die Verfahren zur Festlegung von Grenzwerten und zur Bewertung durch Scores sind.

BEISPIEL 3 Die Planung eines Eignungsprüfungsprogramms erfordert für den Vergleich eines Teilnehmerergebnisses mit den kombinierten Ergebnissen einer Gruppe in derselben Runde und in den Grenzen, die durch die Variabilität von Teilnehmerergebnissen bestimmt wurden, dass die Berechnung eines zugewiesenen Werts und eines geeigneten Maßes der Streuung sowie das Verfahren zur Bewertung durch Scores berücksichtigt werden;

BEISPIEL 4 Die Planung eines Eignungsprüfungsprogramms muss für den Vergleich eines Teilnehmerergebnisses mit dem zugewiesenen Wert unter Verwendung der teilnehmereigenen Messunsicherheit berücksichtigen, wie der zugewiesene Wert und seine Unsicherheit zu erzielen sind und wie teilnehmereigene Messunsicherheiten bei der Bewertung durch Scores zu verwenden sind.

BEISPIEL 5 Die Planung eines Eignungsprüfungsprogramms mit dem Ziel, die Leistung verschiedener Messverfahren zu vergleichen, muss die relevanten zusammenfassenden Kenngrößen und ihre Berechnungsverfahren berücksichtigen.

5.2.2 Es gibt verschiedene Datenarten, die in Eignungsprüfungen verwendet werden, einschließlich quantitative, nominale und ordinale. Unter den quantitativen Variablen könnten einige Ergebnisse auf einer Intervallskala oder einer relativen Skala oder einer Verhältnisskala liegen. Bei einigen Messungen auf einer quantitativen Skala kann nur ein diskreter und unzusammenhängender Satz von Werten realisiert werden (z. B. sequentielle Verdünnungen); in vielen Fällen können diese Ergebnisse jedoch durch Verfahren behandelt werden, die auf stetige quantitative Variablen anwendbar sind.

ANMERKUNG 1 Bei quantitativen Werten ist eine Intervallskala eine Skala, auf der Intervalle (Differenzen) aussagekräftig sind, Verhältnisse jedoch nicht, wie die Celsius-Temperaturskala. Eine Verhältnisskala ist eine Skala, auf der Intervalle und Verhältnisse von Bedeutung sind, wie die Kelvin Temperaturskala oder die üblichsten Längeneinheiten.

ANMERKUNG 2 Bei qualitativen Werten besitzt eine Nominalskala eindeutige Werte, für die die Rangreihung nicht von Bedeutung ist, wie bei Namen von Bakterienarten. Werte auf einer Ordinalskala haben eine aussagekräftige Rangreihung, jedoch sind Differenzen nicht aussagekräftig; z. B. kann eine Skala wie „groß, mittelgroß, klein“ geordnet werden, aber die Differenzen zwischen den Werten sind nicht definiert, anders als bei der Anzahl der Zwischenwerte.

5.2.3 Zusätzlich zu den vorstehend genannten Zwecken dürfen Eignungsprüfungsprogramme für andere Zwecke verwendet werden, wie in 0.1 dargelegt und in ISO/IEC 17043. Die Planung muss für alle angegebenen Zwecke für das bestimmte Eignungsprüfungsprogramm geeignet sein.

5.3 Überlegungen zur statistischen Verteilung der Ergebnisse

5.3.1 Nach ISO/IEC 17043:2010, 4.4.4.2, ist es erforderlich, dass statistische Analyseverfahren mit den statistischen Annahmen für die Daten übereinstimmen. Die gängigsten Analyseverfahren für die Eignungsprüfung gehen davon aus, dass ein Satz von Ergebnissen von kompetenten Teilnehmern annähernd normalverteilt sein wird oder zumindest unimodal und einigermaßen symmetrisch (nach einer Transformation, sofern notwendig) sein wird. Eine gängige zusätzliche Annahme ist, dass die Verteilung der Ergebnisse aus kompetent bestimmten Messungen mit Ergebnissen aus einer Grundgesamtheit von abweichenden Werten, die Ausreißer erzeugen können, gemischt (oder „kontaminiert“) ist. Üblicherweise beruht die Interpretation der Bewertung durch Scores auf der Annahme der Normalität, jedoch nur für die zugrundeliegende angenommene Verteilung bei kompetenten Teilnehmern.

5.3.1.1 Es ist üblicherweise nicht erforderlich zu verifizieren, dass die Ergebnisse normalverteilt sind, aber es ist wichtig eine annähernde Symmetrie zu verifizieren, zumindest visuell. Falls keine Symmetrie verifiziert werden kann, sollte der Anbieter von Eignungsprüfungen Verfahren anwenden, die gegenüber Asymmetrie robust sind (siehe Anhang C).

5.3.1.2 Wenn die für das Eignungsprüfungsprogramm erwartete Verteilung nicht ausreichend symmetrisch ist (mit Berücksichtigung der Kontamination durch Ausreißer), sollte der Anbieter von Eignungsprüfungen Datenanalyseverfahren wählen, die der erwarteten Asymmetrie angemessen Rechnung tragen und die gegen Ausreißer resistent sind, sowie Verfahren zur Bewertung durch Scores, die die erwartete Verteilung für Ergebnisse kompetenter Teilnehmer ebenfalls angemessen berücksichtigen. Das kann umfassen:

- Transformation, um annähernde Symmetrie zu bieten;
- Schätzverfahren, die gegen Asymmetrie resistent sind;
- Schätzverfahren, die geeignete verteilungsbezogene Annahmen umfassen (z. B. maximale Wahrscheinlichkeit einer Übereinstimmung mit geeigneten Verteilungsannahmen und, falls erforderlich, Rückweisung von Ausreißern).

BEISPIEL 1 Ergebnisse, die auf Verdünnung beruhen, wie bei quantitativen mikrobiologischen Zählungen oder bei Immunoassay-Verfahren, sind häufig entsprechend der logarithmischen Normalverteilung verteilt, und daher kann eine logarithmische Transformation als erster Analyseschritt geeignet sein.

BEISPIEL 2 Zählungen von kleinen Partikelanzahlen können entsprechend einer Poisson-Verteilung verteilt sein; und deshalb können die Kriterien für die Leistungsbewertung anhand einer Tabelle von Poisson-Wahrscheinlichkeiten bestimmt werden, auf Grundlage des Mittelwerts der Zählung für die Teilnehmergruppe.

5.3.1.3 In einigen Bereichen der Kalibrierung können Teilnehmerergebnisse statistischen Verteilungen folgen, die im Messverfahren beschrieben sind (z. B. exponentiell oder wellenförmig); diese festgelegten Verteilungen sollten in jedem Bewertungsprotokoll berücksichtigt werden.

5.3.2 In Übereinstimmung mit ISO/IEC 17043:2010, 4.4.4.2, muss der Anbieter von Eignungsprüfungen die Grundlage jeglicher statistischen Annahmen angeben und nachweisen, dass sie begründet sind. Dieser Nachweis kann z. B. auf Grundlage der beobachteten Daten, von Ergebnissen aus früheren Runden des Eignungsprüfungsprogramms oder der technischen Fachliteratur erfolgen.

ANMERKUNG Der Nachweis der Begründbarkeit von Verteilungsannahmen ist weniger streng als der Nachweis der Gültigkeit dieser Annahme.

5.4 Überlegungen für kleine Teilnehmeranzahlen

5.4.1 Die statistische Planung eines Eignungsprüfungsprogramms muss die Mindestanzahl der Teilnehmer berücksichtigen, die erforderlich ist, um die Ziele der Planung zu erfüllen und alternative Ansätze angeben, die angewendet werden, falls die Mindestanzahl nicht erreicht wird (ISO/IEC 17043:2010, 4.4.4.3 b)). Statistische Verfahren, die für eine große Teilnehmeranzahl geeignet sind, können für begrenzte Teilnehmeranzahlen nicht geeignet sein. Es gibt Bedenken, dass Kenngrößen, die aus einer kleinen Anzahl von Teilnehmerergebnissen bestimmt wurden, nicht ausreichend zuverlässig sein können und ein Teilnehmer gegen eine ungeeignete Vergleichsgruppe bewertet werden könnte.

ANMERKUNG Der Technische Bericht der IUPAC/CITAC: *Selection and use of proficiency testing schemes for a limited number of participants* [24] enthält einen nützlichen Leitfaden für Eignungsprüfungsprogramme mit wenigen Teilnehmern. Kurz gesagt: der IUPAC/CITAC-Bericht empfiehlt, dass der zugewiesene Wert auf zuverlässigen unabhängigen Messungen beruhen sollte, z. B. durch Verwendung von zertifiziertem Referenzmaterial, unabhängige Zuweisung durch eine Kalibrier- oder staatliche metrologische Stelle oder durch gravimetrische Vorbereitung. Der Bericht gibt ferner an, dass die Standardabweichung für die Eignungsbeurteilung nicht auf der beobachteten Streuung unter den Teilnehmerergebnissen bei einer einzelnen Runde des Eignungsprüfungsprogramms beruhen sollte.

5.4.2 Die Mindestanzahl von Teilnehmern, die für die verschiedenen statistischen Verfahren erforderlich sind, wird von einer Vielzahl von Situationen abhängig sein:

- den angewendeten statistischen Verfahren, z. B. dem gewählten, jeweiligen robusten Verfahren oder der gewählten Strategie zum Entfernen der Ausreißer;
- der Erfahrung der Teilnehmer mit dem speziellen Eignungsprüfungsprogramm;
- der Erfahrung des Anbieters von Eignungsprüfungen mit der Matrix, Messgröße, Verfahren und Teilnehmergruppe;
- ob es beabsichtigt ist, den zugewiesenen Wert oder die Standardabweichung (oder beide) zu bestimmen.

Ein weiterführender Leitfaden zu Verfahren zur Behandlung einer kleinen Gruppe von Teilnehmern ist in Anhang D.1 enthalten.

5.5 Leitlinien zur Auswahl des Berichtsformats

5.5.1 Es ist eine Anforderung nach ISO/IEC 17043:2010, 4.6.1.2, dass Anbieter von Eignungsprüfungen die Teilnehmer dahingehend instruieren, dass diese die Messungen so durchführen und die Ergebnisse zu Prüfgegenständen für die Eignungsprüfung so berichten, wie sie das für die Mehrheit der routinemäßig durchgeführten Messungen tun, ausgenommen bei besonderen Umständen.

Diese Anforderung kann in einigen Fällen erschweren, eine genaue Beurteilung eines Teilnehmers hinsichtlich seiner Präzision und Richtigkeit oder Kompetenz für ein Messverfahren zu erzielen. Der Anbieter von Eignungsprüfungen sollte ein konsistentes Berichterstattungsformat für das Eignungsprüfungsprogramm festlegen, aber er sollte wenn möglich Einheiten verwenden, die der Mehrheit der Teilnehmer vertraut sind und ein Berichterstattungsformat wählen, dass Übertragungsfehler und andere Fehler auf ein Mindestmaß beschränkt. Das kann eine automatisierte Warnung vor ungeeigneten Einheiten einschließen, falls bekannt ist, dass Teilnehmer routinemäßig in anderen Einheiten berichten, als in den durch das Programm geforderten.

ANMERKUNG 1 Bei einigen Eignungsprüfungsprogrammen ist es Ziel, die Fähigkeit eines Teilnehmers zur Einhaltung eines Standardverfahrens zu bewerten, was die Verwendung einer bestimmten Maßeinheit oder Anzahl signifikanter Stellen einschließen könnte.

ANMERKUNG 2 Übertragungsfehler beim Zusammenführen von Ergebnissen durch den Anbieter von Eignungsprüfungen können durch Verwendung elektronischer Berichterstattungssysteme erheblich verringert oder ausgeschlossen werden, die Teilnehmern ermöglichen, ihre eigenen Daten direkt einzugeben.

5.5.2 Falls ein Eignungsprüfungsprogramm Wiederholmessungen an Prüfgegenständen für die Eignungsprüfung erfordert, sollte der Teilnehmer aufgefordert werden, alle Wiederholungsprüfwerte zu berichten. Das kann z. B. eintreten, wenn ein Ziel darin besteht, die Präzision eines Teilnehmers zu bewerten hinsichtlich Prüfgegenstände, die bekanntermaßen Wiederholproben sind oder wenn ein Messverfahren eine gesonderte Berichterstattung von Mehrfachbeobachtungen erfordert. In diesen Fällen kann es erforderlich sein, dass der Anbieter von Eignungsprüfungen auch den Mittelwert des Teilnehmers (oder einen anderen Schätzwert der Lage) und die Unsicherheit abfragt, um die Datenanalyse durch den Anbieter von Eignungsprüfungen zu unterstützen.

5.5.3 Wenn die konventionelle Praxis der Berichterstattung darin besteht, Ergebnisse so zu berichten, dass sie „kleiner als“ oder „größer als“ eine Grenze (z. B. ein Kalibrierniveau oder eine Quantifizierungsgrenze) sind, und wenn numerische Ergebnisse für die Bewertung durch Scores erforderlich sind, muss der Anbieter von Eignungsprüfungen bestimmen, wie die Ergebnisse verarbeitet werden.

5.5.3.1 Der Anbieter von Eignungsprüfungen sollte entweder validierte Verfahren zur Datenbehandlung und zur Bewertung durch Scores festlegen, die zensierte Daten^{N1)} enthalten (siehe Anhang E.1) oder von Teilnehmern fordern, dass sie den Messwert des Ergebnisses entweder anstelle oder zusätzlich zum konventionell berichteten Wert berichten.

ANMERKUNG 1 Eine Möglichkeit des Verfahrens zur Bewertung durch Scores könnte sein, derartige Daten nicht zu bewerten.

ANMERKUNG 2 Wenn von Teilnehmern gefordert wird, numerische Werte zu berichten, die außerhalb der üblicherweise berichteten Spannweite liegen (z. B. unterhalb der Bestimmungsgrenze des Teilnehmers), wird dies die Anwendung statistischer Verfahren ermöglichen, die numerische Werte erfordern, es kann jedoch zu Werten führen, die nicht die übliche Dienstleistung widerspiegeln, die der Teilnehmer an Kunden erbringt.

N1) Nationale Fußnote: Unter „zensierte Daten“ werden Werte verstanden, die z. B. als „< x“ angegeben werden.

5.5.3.2 Wenn Konsenskenngößen verwendet werden, kann die Bewertung der Leistung nicht möglich sein, wenn die Anzahl der zensierten Werte groß genug ist, dass ein robustes Verfahren durch das Zensieren beeinträchtigt wird. Falls die Anzahl zensierter Ergebnisse ausreicht, um ein robustes Verfahren zu beeinflussen, sollten die Ergebnisse mithilfe statistischer Verfahren bewertet werden, die eine erwartungstreue Schätzung bei Vorliegen von zensierten Messreihen [21] ermöglichen oder die Ergebnisse sollten nicht bewertet werden. Wenn Zweifel über die Wirkung des gewählten Verfahrens bestehen, sollte der Anbieter von Eignungsprüfungen zusammenfassende Kenngrößen und Leistungsbewertungen mit jedem der alternativen statistischen Verfahren berechnen, die unter den Umständen als potentiell anwendbar angesehen werden und die Bedeutung etwaiger Differenz(en) untersuchen.

5.5.3.3 Wenn zensierte Ergebnisse, wie Angaben „kleiner als“, erwartet werden oder beobachtet wurden, sollte die Planung des Eignungsprüfungsprogramms Vorkehrungen für eine Bewertung durch Scores und/oder andere Handlungen an von Teilnehmern berichteten zensierten Werten enthalten, und die Teilnehmer sollten über diese Vorkehrungen unterrichtet werden.

ANMERKUNG Anhang E.1 enthält ein Beispiel für einige Analyseansätze für zensierte Messreihen. Dieses Beispiel zeigt robuste Konsenskenngößen mit drei verschiedenen Ansätzen; wobei die Werte aus der zensierten Messreihe entfernt werden, wobei die Werte beibehalten werden, aber das „<“-Zeichen entfernt wird, und wobei die Ergebnisse durch die Hälfte des Grenzwerts ersetzt werden.

5.5.4 Üblicherweise wird die Anzahl der zu berichtenden signifikanten Stellen durch die Planung des Eignungsprüfungsprogramms bestimmt.

5.5.4.1 Bei der Festlegung der zu berichtenden Anzahl der signifikanten Stellen, sollten die Rundungsfehler im Vergleich mit der erwarteten Streuung zwischen den Teilnehmern vernachlässigbar sein.

ANMERKUNG In einigen Fällen gehört die korrekte Berichterstattung zur Kompetenzbestimmung des Teilnehmers und die Anzahl der signifikanten Stellen und Dezimalstellen kann variieren.

5.5.4.2 Wenn die Anzahl von Dezimalstellen, die unter Bedingungen einer Routinemessung berichtet werden, eine merkliche nachteilige Auswirkung auf die Datenbehandlung durch den Anbieter von Eignungsprüfungen hat (z.B. wenn Messverfahren die Berichterstattung bis auf eine kleine Anzahl wesentlicher Dezimalstellen erfordern), darf der Anbieter von Eignungsprüfungen die Anzahl der zu berichtenden Dezimalstellen festlegen.

BEISPIEL Ein Messverfahren könnte die Berichterstattung bis auf 0,1 g festlegen, was zu einem großen Anteil (> 50 %) von identischen Ergebnissen führt und im Gegenzug die Berechnung von robusten Mittelwerten und Standardabweichungen beeinträchtigt. Der Anbieter von Eignungsprüfungen könnte dann von den Teilnehmern fordern, bis zu zwei oder drei Dezimalstellen zu berichten, um ausreichend zuverlässige Schätzwerte der Lage und Streuung zu erhalten.

5.5.4.3 Wenn es zulässig ist, dass verschiedene Teilnehmer ihre Ergebnisse mit unterschiedlichen Anzahlen an wesentlichen Dezimalstellen berichten, sollte das der Anbieter von Eignungsprüfungen bei der Erstellung von Konsenskenngößen (z.B. zugewiesener Wert und Standardabweichung für die Eignungsbeurteilung) berücksichtigen.

6 Leitlinien für die anfängliche Überprüfung von Prüfgegenständen für die Eignungsprüfungen und von Ergebnissen

6.1 Homogenität und Stabilität von Prüfgegenständen für die Eignungsprüfung

6.1.1 Der Anbieter von Eignungsprüfungen muss sicherstellen, dass die Serie von Prüfgegenständen ausreichend homogen und stabil für die Zwecke des Eignungsprüfungsprogramms sind. Der Anbieter muss die Homogenität und Stabilität beurteilen, indem er Kriterien anwendet, die sicherstellen, dass Inhomogenität und Instabilität der Prüfgegenstände die Leistungsbewertung nicht nachteilig beeinflussen. Für die Beurteilung der Homogenität und Stabilität sollte einer oder mehrere der folgenden Ansätze verwendet werden:

- a) experimentelle Untersuchungen wie nach Anhang B oder alternative experimentelle Verfahren, mit denen die Homogenität und Stabilität in gleichwertigem oder in größerem Maße sichergestellt wird;
- b) Erfahrung mit dem Verhalten sehr ähnlicher Prüfgegenstände für die Eignungsprüfung in früheren Runden des Eignungsprüfungsprogramms, mit der notwendigen Verifizierung in der aktuellen Runde.
- c) Beurteilung der Teilnehmerdaten in der aktuellen Runde des Eignungsprüfungsprogramms hinsichtlich des Nachweises der Konsistenz mit früheren Runden, hinsichtlich des Nachweises von Änderung mit der Berichtszeit oder Herstellungsreihenfolge oder jeglicher unerwarteter Streuung, die Inhomogenität oder Instabilität zugeschrieben werden können.

ANMERKUNG 1 Diese Ansätze können fallweise übernommen werden, indem geeignete statistische Verfahren und technische Begründung angewendet werden. Der Ansatz wird sich häufig während der Lebensdauer eines Eignungsprüfungsprogramms ändern, z. B. wenn gesammelte Erfahrung die anfängliche Anforderung nach experimenteller Untersuchung verringert.

ANMERKUNG 2 Es ist nur solange vernünftig, sich auf Erfahrung zu verlassen (wie im vorstehenden Listenelement b), wie:

- 1) sich der Prozess der Herstellung von Serien von Prüfgegenständen für die Eignungsprüfung nicht in einer Weise ändert, die die Homogenität beeinflussen kann;
- 2) sich die bei der Herstellung der Prüfgegenstände für die Eignungsprüfung verwendeten Materialien nicht in einer Weise ändern, die die Homogenität beeinflussen kann;
- 3) kein „Fehler“ bei der Homogenität erkannt wird, entweder durch eine Homogenitätsprüfung oder durch Reaktionen von Teilnehmern; und
- 4) die Homogenitätsanforderungen an das Material regelmäßig überprüft werden, unter Berücksichtigung des vorgesehenen Verwendungszwecks des Materials zum Zeitpunkt der Überprüfung, um sicherzustellen, dass die durch den Herstellungsprozess erreichte Homogenität für den Verwendungszweck geeignet bleibt.

BEISPIEL Der Nachweis ausreichender Homogenität und Stabilität in der aktuellen Runde ist erbracht, wenn in früheren Runden eines Eignungsprüfungsprogramms Prüfgegenstände für die Eignungsprüfung verwendet wurden, die geprüft und deren ausreichende Homogenität und Stabilität nachgewiesen wurden und wenn die Teilnehmer dieselben wie in früheren Runden sind, und wenn die Standardabweichung zwischen Laboratorien in der aktuellen Runde die Standardabweichung früherer Runden nicht überschreitet.

6.1.2 Bei Eignungsprüfungsprogrammen für Kalibrierungen, bei denen derselbe Gegenstand von mehreren Teilnehmern verwendet wird, muss der Anbieter von Eignungsprüfungen die Stabilität über die gesamte Runde sicherstellen oder über Verfahren verfügen, um Instabilität im Fortgang einer Runde des Eignungsprüfungsprogramms zu erkennen und ihr Rechnung zu tragen. Das sollte einschließen, dass Tendenzen bei bestimmten Prüfgegenständen und Messgrößen, z. B. Drift, berücksichtigt werden. Für die Sicherstellung der Stabilität sollten, soweit zutreffend, die Auswirkungen des mehrfachen Versands desselben Gegenstands berücksichtigt werden.

6.1.3 Sämtliche Messgrößen (oder Eigenschaften) sollten üblicherweise auf Homogenität und Stabilität überprüft werden. Wenn das Verhalten einer Teilmenge von Eigenschaften nachweisbar einen guten Hinweis auf die Stabilität und/oder Homogenität sämtlicher Eigenschaften gibt, die in einer Runde berichtet werden, darf die Beurteilung nach 6.1.1 auf diese Teilmenge an Eigenschaften beschränkt werden. Die überprüften Messgrößen sollten empfindlich auf Quellen von Inhomogenität oder Instabilität bei der Verarbeitung des Prüfgegenstands für die Eignungsprüfung reagieren. Einige wichtige Fälle sind:

- a) wenn ein Anteil zu messen ist, kann ein kleiner Anteil als Merkmal schwieriger zu homogenisieren sein und daher bei einer Homogenitätsprüfung empfindlicher reagieren;
- b) falls ein Prüfgegenstand für die Eignungsprüfung während der Verarbeitung erwärmt wird, wird eine Messgröße gewählt, die empfindlich auf ungleichmäßige Erwärmung reagiert;
- c) falls eine gemessene Eigenschaft durch Absetzen, Niederschlag oder andere zeitabhängige Wirkungen während der Vorbereitung von Prüfgegenständen für die Eignungsprüfung beeinflusst werden kann, sollte diese Eigenschaft übergreifend über die Reihenfolge des Befüllens überprüft werden.

BEISPIEL In einem Eignungsprüfungsprogramm für den Gehalt an toxischen Metallen von Böden wird der gemessene Metallgehalt hauptsächlich durch den Feuchtegehalt beeinflusst. Dadurch kann eine Überprüfung auf einen konsistenten Feuchtegehalt für die Sicherstellung einer angemessenen Stabilität toxischer Metalle als ausreichend angesehen werden.

ANMERKUNG Ein Beispiel für Homogenitäts- und Stabilitätsprüfungen ist in Anhang E.2 enthalten, unter Anwendung statistischer Verfahren, die in Anhang B empfohlen werden.

6.2 Betrachtungen für verschiedene Messverfahren

6.2.1 Wenn von allen Teilnehmern erwartet wird, dass sie einen Wert für dieselbe Messgröße berichten, sollte der zugewiesene Wert üblicherweise bei allen Teilnehmern derselbe sein. Wird den Teilnehmern jedoch gestattet, ihr Messverfahren selbst zu wählen, ist es möglich, dass ein einzelner zugewiesener Wert für jeden Analyt oder jede Eigenschaft nicht für alle Teilnehmer geeignet ist. Das kann z. B. auftreten, wenn unterschiedliche Messverfahren Ergebnisse liefern, die nicht vergleichbar sind. In diesem Fall darf der Anbieter von Eignungsprüfungen für jedes Messverfahren einen unterschiedlichen zugewiesenen Wert verwenden.

BEISPIELE:

- a) medizinische Prüfungen, bei denen für unterschiedliche zugelassene Messverfahren bekannt ist, dass sie unterschiedlich auf dasselbe Prüfmaterial reagieren und verschiedene Referenzbereiche für die Diagnose verwenden;
- b) operativ festgelegte Messgrößen, wie herauslösbare giftige Metalle in Böden, für die andere Standardverfahren verfügbar sind und von denen nicht erwartet wird, dass sie direkt verglichen werden, bei denen jedoch das Eignungsprüfungsprogramm die Messgröße ohne Bezug auf ein spezifisches Prüfverfahren festlegt.

6.2.2 Der Bedarf für verschiedene zugewiesene Werte für Teilmengen von Teilnehmern sollte in der Planung des Eignungsprüfungsprogramms berücksichtigt werden (z. B. um für die Berichterstattung spezieller Verfahren vorzusorgen) und sollte auch bei der Datenüberprüfung für jede Runde berücksichtigt werden.

6.3 Entfernung grober Fehler

6.3.1 ISO/IEC 17043:2010, B.2.5 und das IUPAC Harmonized Protocol empfehlen das Entfernen offensichtlicher grober Fehler aus dem Datensatz in einer frühen Phase einer Analyse, vor der Anwendung jeglicher robuster Verfahren oder jeglicher Prüfung zur Identifizierung statistischer Ausreißer. Im Allgemeinen würden diese Ergebnisse gesondert behandelt werden (wie das Kontaktieren eines Teilnehmers). Es kann möglich sein, einige grobe Fehler zu korrigieren, das sollte jedoch ausschließlich in Übereinstimmung mit einer bewährten Politik und einem bewährten Verfahren ausgeführt werden.

ANMERKUNG Offensichtliche grobe Fehler, wie das Berichten von Ergebnissen in falschen Einheiten oder das Vertauschen von Ergebnissen unterschiedlicher Prüfgegenstände für die Eignungsprüfung treten in den meisten Runden von Eignungsprüfungen auf, und diese Ergebnisse beeinträchtigen nur die Leistung der nachfolgenden statistischen Verfahren.

6.3.2 Falls Zweifel bestehen, ob ein Ergebnis einen groben Fehler darstellt, sollte es im Datensatz beibehalten und nachfolgender Behandlung unterzogen werden, wie nach 6.4 bis 6.6 beschrieben.

6.4 Visuelle Überprüfung von Daten

6.4.1 Als einen ersten Schritt in jeder Datenanalyse sollte der Anbieter die visuelle Überprüfung der Daten vorsehen, die durch eine Person mit einer angemessenen technischen und statistischen Fachkenntnis durchgeführt wird. Diese Überprüfung dient der Bestätigung der erwarteten Verteilung der Ergebnisse und dem Erkennen von Anomalien oder unvorhergesehenen Quellen von Variabilität. Zum Beispiel könnte eine bimodale Verteilung ein Hinweis auf eine gemischte Grundgesamtheit von Ergebnissen sein, verursacht durch unterschiedliche Verfahren, kontaminierte Proben oder schlecht formulierte Anweisungen. In dieser Situation sollten solche Bedenken ausgeräumt werden, bevor die Auswertung oder Beurteilung fortgesetzt wird.

ANMERKUNG 1 Ein Histogramm ist ein nützliches und allgemein verfügbares Überprüfungsverfahren, um nach einer Verteilung zu suchen, die unimodal und symmetrisch ist, und um unübliche Ausreißer (10.2) zu erkennen. Die Intervalle, die dazu verwendet werden, die Ergebnisse in einem Histogramm zusammenzuführen, reagieren jedoch empfindlich auf die Anzahl der Ergebnisse und Schnittpunkte und können deshalb schwierig zu erstellen sein. Häufig ist ein Kerndichtediagramm für das Erkennen möglicher Bimodalitäten oder fehlender Symmetrie (10.3) von größerem Nutzen.

ANMERKUNG 2 Andere Überprüfungsverfahren können nützlich sein, wie ein kumulatives Verteilungsdiagramm oder ein Stamm-Blatt-Diagramm. Einige graphische Verfahren für die Datenüberprüfung werden in Anhang E.3 und E.4 dargestellt.

6.4.2 Wenn die Durchführung einer visuellen Überprüfung aller interessierenden Datensätze nicht realisierbar ist, muss es ein Verfahren zur Warnung vor unerwarteter Variabilität in einem Datensatz geben, z. B. durch Überprüfung der Unsicherheit des zugewiesenen Werts im Vergleich mit den Bewertungskriterien oder durch Vergleich mit früheren Runden des Eignungsprüfungsprogramms.

6.5 Robuste statistische Verfahren

6.5.1 Robuste statistische Verfahren können angewendet werden, um den zentralen Teil eines normalverteilten Satzes von Ergebnissen zu beschreiben, jedoch ohne dass das Erkennen spezifischer Werte als Ausreißer und deren Ausschluss aus nachfolgenden Analysen erforderlich ist. Viele angewendete robuste Verfahren beruhen (im ersten Schritt) auf dem Median und der Spannweite der zentralen 50 % der Ergebnisse; diese sind Maße der Lage und der Streuung der Daten, die dem Mittelwert und der Standardabweichung ähneln. Im Allgemeinen sollten robuste Verfahren den Verfahren vorgezogen werden, die als Ausreißer gekennzeichnete Ergebnisse löschen.

ANMERKUNG Strategien, die klassische Kenngrößen verwenden, wie die Standardabweichung nach dem Entfernen von Ausreißern, führen üblicherweise zu Unterschätzungen der Streuung bei annähernd normal verteilten Daten; robuste Kenngrößen werden üblicherweise angepasst, um erwartungstreue Schätzwerte der Streuung zu erhalten.

6.5.2 Der Median, der skalierte Median der absoluten Abweichungen (*MAD_e*) und der normierte IQR (*nIQR*) sind als einfache Schätzfunktionen zulässig. Der Algorithmus A wandelt die ursprünglichen Daten durch einen Prozess um, der als Winsorisieren bezeichnet wird, um alternative Schätzer des Mittelwerts und der Standardabweichung für annähernd normal verteilte Daten bereitzustellen und ist am nützlichsten, wenn der erwartete Anteil der Ausreißer unter 20 % liegt. Die Q_n - und Q -Verfahren (nach Anhang C) für die Schätzung der Standardabweichung sind insbesondere in Fällen nützlich, in denen ein großer Anteil (> 20 %) der Ergebnisse abweichen kann oder in denen Daten nicht zuverlässig durch Experten überprüft werden können. Weitere in Anhang C beschriebene Verfahren bieten ebenfalls gute Leistung, wenn der erwartete Anteil an Extremwerten über 20 % liegt (siehe Anhang D).

ANMERKUNG Der Median, der Interquartilsabstand und der skalierte Median der absoluten Abweichungen haben eine größere Streuung als der Mittelwert und die Standardabweichung, wenn sie auf annähernd normalverteilte Daten angewendet werden. Komplexere robuste Schätzer bieten eine bessere Leistung bei annähernd normalverteilten Daten, wobei sie viel von der Resistenz gegen Ausreißerergebnisse bewahren, die durch den Median und den Interquartilsabstand geboten wird.

6.5.3 Die Auswahl der statistischen Verfahren obliegt dem Anbieter von Eignungsprüfungen. Der robuste Mittelwert und die robuste Standardabweichung können für verschiedene Zwecke verwendet werden, von denen die Leistungsbewertung nur einer ist. Robuste Mittelwerte und Standardabweichungen können auch als zusammenfassende Kenngrößen für verschiedene Gruppen von Teilnehmern oder für besondere Verfahren verwendet werden.

ANMERKUNG Einzelheiten zu robusten Verfahren sind in Anhang C enthalten. Anhang E.3 und E.4 enthalten umfangreiche Beispiele, die die Anwendung einer Vielzahl robuster statistischer Verfahren nach Anhang C veranschaulichen.

6.6 Ausreißer Techniken für Einzelergebnisse

6.6.1 Ausreißerprüfungen können angewendet werden, um entweder die visuelle Überprüfung auf Anomalien zu unterstützen oder, zusammen mit der Rückweisung von Ausreißern, bei der Berechnung zusammenfassender Kenngrößen einen Grad an Resistenz gegen Extremwerte zu bieten. Wenn Verfahren zur Ausreißererkennung angewendet werden, sollte nachgewiesen werden, dass die der Prüfung zugrundeliegenden Annahmen in ausreichendem Maß dem Zweck des Eignungsprüfungsprogramms entsprechen; insbesondere gehen viele Ausreißerprüfungen von einer zugrundeliegenden Normalverteilung aus.

ANMERKUNG ISO 16269-4 [10] und ISO 5725-2 [1] enthalten mehrere Verfahren zur Ausreißererkennung, die auf Ringversuchsdaten anwendbar sind.

6.6.2 Strategien zur Rückweisung von Ausreißern, die darauf beruhen, dass Ausreißer zurückgewiesen werden, die durch eine Ausreißerprüfung mit einem sehr hohen Vertrauensniveau erkannt wurden, gefolgt von der Anwendung einfacher Kenngrößen, wie dem Mittelwert und der Standardabweichung, sind zulässig, wenn robuste Verfahren nicht anwendbar sind (siehe 6.5.1). Wenn Strategien zur Rückweisung von Ausreißern angewendet werden, muss der Anbieter von Eignungsprüfungen:

- a) die Prüfungen und das Vertrauensniveau, die für die Rückweisung erforderlich sind dokumentieren;
- b) Grenzwerte für den Anteil der durch aufeinanderfolgende Ausreißerprüfungen, falls zutreffend, rückgewiesenen Daten festlegen;
- c) nachweisen, dass die sich daraus ergebenden Schätzwerte von Lage und (sofern zutreffend) Skala eine für den Zweck des Eignungsprüfungsprogramms ausreichende Leistung haben (einschließlich Effizienz und systematischer Abweichung).

ANMERKUNG ISO 5725-2 enthält Empfehlungen für das geeignete Vertrauensniveau für die Rückweisung von Ausreißern in Ringversuchen für die Bestimmung der Präzision von Prüfverfahren. Insbesondere ISO 5725-2 empfiehlt die Rückweisung nur bei dem Niveau von 99 %, es sei denn, es liegt ein anderer guter Grund vor, um ein bestimmtes Ergebnis rückzuweisen.

6.6.3 Wenn die Rückweisung von Ausreißern Teil des Datenbehandlungsverfahrens ist und ein Ergebnis als Ausreißer entfernt wird, sollte die Teilnehmerleistung weiterhin bewertet werden, in Übereinstimmung mit den Kriterien, die auf alle Teilnehmer im Eignungsprüfungsprogramm angewendet werden.

ANMERKUNG 1 Ausreißer unter den berichteten Werten werden häufig unter Anwendung der Grubbs-Ausreißerprüfung nach ISO 5725-2 identifiziert. In diesem Verfahren wird die Bewertung unter Anwendung der Standardabweichung aller Teilnehmer einschließlich potentieller Ausreißer angewendet. Daher sollte dieses Verfahren angewendet werden, wenn die Leistung der Teilnehmer den Erwartungen aus früheren Runden entspricht und es eine kleine Menge an Ausreißern gibt (ein oder zwei Ausreißer auf jeder Seite des Mittelwerts). Herkömmliche Tabellen für das Grubbsverfahren setzen eine einzelne Anwendung für einen möglichen Ausreißer (oder zwei) an einem festgelegten Ort, nicht unbegrenzte sequenzielle Anwendung voraus. Falls die Grubbs-Tabellen sequentiell verwendet werden, können die Wahrscheinlichkeiten des Fehlers 1. Art nicht gelten.

ANMERKUNG 2 Wenn Ergebnisse von Wiederholmessungen berichtet werden oder identische Prüfgegenstände für die Eignungsprüfung in einer Runde des Eignungsprüfungsprogramms verwendet werden, wird üblicherweise die Cochran-Prüfung für Ausreißer bei der Wiederholpräzision angewendet, auch beschrieben in ISO 5725-2.

ANMERKUNG 3 Ausreißer können auch durch robuste oder nichtparametrische Verfahren identifiziert werden; z. B. bei der Berechnung eines robusten Mittelwerts und einer robusten Standardabweichung könnten Werte, die vom robusten Mittelwert um mehr als das Dreifache der robusten Standardabweichung abweichen, als Ausreißer identifiziert werden.

7 Bestimmung des zugewiesenen Werts und seiner Standardunsicherheit

7.1 Auswahl des Verfahrens zur Bestimmung des zugewiesenen Werts

7.1.1 In 7.3 bis 7.7 werden fünf Arten zur Bestimmung des zugewiesenen Werts x_{pt} beschrieben. Die Auswahl zwischen diesen Verfahren obliegt der Verantwortung des Anbieters von Eignungsprüfungen.

ANMERKUNG Die Ansätze nach 7.3 bis 7.6 weisen eine große Ähnlichkeit zu den Ansätzen auf, die für die Bestimmung der Eigenschaftswerte von zertifizierten Referenzmaterialien nach ISO Guide 35 [13] verwendet werden.

7.1.2 Alternative Verfahren zur Bestimmung des zugewiesenen Werts und seiner Unsicherheit dürfen angewendet werden, vorausgesetzt, dass sie eine fundierte statistische Grundlage aufweisen und dass das angewendete Verfahren im dokumentierten Plan für das Eignungsprüfungsprogramm beschrieben ist und den Teilnehmern vollständig beschrieben wird. Ungeachtet dessen, welches Verfahren zur Bestimmung des zugewiesenen Werts angewendet wird, ist es immer angebracht, die Gültigkeit des zugewiesenen Werts für diese Runde eines Eignungsprüfungsprogramms zu überprüfen. Das wird in 7.8 dargelegt.

7.1.3 Ansätze zur Bestimmung qualitativer zugewiesener Werte werden in 11.3 ausgeführt.

7.1.4 Das Verfahren zur Bestimmung des zugewiesenen Werts und seiner zugehörigen Unsicherheit muss in jedem Bericht an die Teilnehmer angegeben werden oder in einem Programmprotokoll, das allen Teilnehmern zur Verfügung steht, deutlich beschrieben werden.

7.2 Bestimmung der Unsicherheit des zugewiesenen Werts

7.2.1 Der *Guide to the expression of uncertainty in measurement* (ISO/IEC Guide 98-3 [14]) enthält einen Leitfaden zur Berechnung von Messunsicherheiten. ISO Guide 35 enthält einen Leitfaden zur Unsicherheit des zugewiesenen Werts für zertifizierte Eigenschaftswerte, der auf viele Planungen für Eignungsprüfungsprogramme angewendet werden kann.

7.2.2 Ein allgemeines Modell für den zugewiesenen Wert und seine Unsicherheit wird in den Gleichungen (2) und (3) beschrieben:

Das Modell für den zugewiesenen Wert kann wie folgt dargestellt werden:

$$x_{pt} = x_{char} + \delta_{hom} + \delta_{trans} + \delta_{stab} \quad (2)$$

Dabei ist

x_{pt}	der zugewiesene Wert;
x_{char}	der Eigenschaftswert, der aus der Charakterisierung (Bestimmung des zugewiesenen Werts) erhalten wurde;
δ_{hom}	ein Abweichungsterm aufgrund der Abweichung zwischen Prüfgegenständen für die Eignungsprüfung;
δ_{trans}	ein Abweichungsterm aufgrund der Instabilität unter Transportbedingungen;
δ_{stab}	ein Abweichungsterm aufgrund der Instabilität während der Dauer der Eignungsprüfung.

Das zugehörige Modell für die Unsicherheit des zugewiesenen Werts kann wie folgt dargestellt werden:

$$u(x_{pt}) = \sqrt{u_{char}^2 + u_{hom}^2 + u_{trans}^2 + u_{stab}^2} \quad (3)$$

Dabei ist

$u(x_{pt})$	die Standardunsicherheit des zugewiesenen Werts;
u_{char}	die Standardunsicherheit aufgrund der Charakterisierung;
u_{hom}	die Standardunsicherheit aufgrund der Abweichungen zwischen Prüfgegenständen für die Eignungsprüfung;
u_{trans}	die Standardunsicherheit aufgrund der Instabilität durch den Transport von Prüfgegenständen für die Eignungsprüfung;
u_{stab}	die Standardunsicherheit aufgrund der Instabilität während der Dauer der Eignungsprüfung;

ANMERKUNG 1 Sowohl Kovarianz zwischen Quellen von Unsicherheit als auch vernachlässigbare Quellen können bei speziellen Anwendungen zu einem anderen Modell führen. In einigen Fällen kann jede der Unsicherheitskomponenten null betragen oder vernachlässigbar sein.

ANMERKUNG 2 Wenn σ_{pt} als die Standardabweichung von Teilnehmerergebnissen berechnet wird, werden die Unsicherheitskomponenten aufgrund von Inhomogenität, Transport und Instabilität zum großen Teil in der Variabilität der Teilnehmerergebnisse widerspiegelt. In diesem Fall ist die Unsicherheit der Charakterisierung, wie in 7.3 bis 7.7 beschrieben, ausreichend.

ANMERKUNG 3 Vom Anbieter von Eignungsprüfungen wird üblicherweise erwartet, dass er sicherstellt, dass Änderungen bezüglich Instabilität oder auf dem Transport entstandene Änderungen im Vergleich mit der Standardabweichung für die Eignungsbeurteilung vernachlässigbar sind; das heißt, sicherzustellen, dass δ_{trans} und δ_{stab} vernachlässigbar sind. Wenn diese Anforderung erfüllt wird, können u_{stab} und u_{trans} auf null gesetzt werden.

7.2.3 Der zugewiesene Wert kann eine systematische Abweichungskomponente enthalten, die in dem vorstehenden Modell nicht berücksichtigt ist. Diese muss, falls möglich, in der Planung des Eignungsprüfungsprogramms berücksichtigt werden. Falls die systematische Abweichungskomponente im zugewiesenen Wert korrigiert wird, muss die Unsicherheit dieser Korrektur in die Bewertung der Unsicherheit des zugewiesenen Werts aufgenommen werden.

7.3 Bekannte Werte aus der Probenherstellung

7.3.1 Der Prüfgegenstand für die Eignungsprüfung kann durch Mischen von Materialien mit bekannten Niveaus einer Eigenschaft in festgelegten Anteilen oder durch das Hinzufügen eines festgelegten Anteils einer Substanz zu einem Ausgangsstoff hergestellt werden.

7.3.1.1 Der zugewiesene Wert x_{pt} wird durch Berechnung aus den Massen der verwendeten Eigenschaften abgeleitet. Dieser Ansatz ist besonders wertvoll, wenn auf diese Weise einzelne Prüfgegenstände für die Eignungsprüfung hergestellt werden; und es sind die Anteile der Eigenschaften zu bestimmen.

7.3.1.2 Es sollte sorgfältig darauf geachtet werden, sicherzustellen, dass:

- a) der Ausgangsstoff effektiv keinen Anteil des hinzugefügten Bestandteils enthält oder dass der Anteil des hinzugefügten Bestandteils im Ausgangsstoff genau bekannt ist;
- b) die Bestandteile homogen vermischt werden (soweit erforderlich);
- c) sämtliche signifikanten Fehlerquellen identifiziert werden (z. B. wird nicht immer erkannt, dass Glas Quecksilberverbindungen absorbiert, so dass die Konzentration einer wässrigen Lösung einer Quecksilberverbindung durch den Behälter verändert werden kann);
- d) keine nachteilige Wechselwirkung zwischen den Bestandteilen und der Probenmatrix stattfindet;
- e) das Verhalten der Prüfgegenstände für die Eignungsprüfung, die hinzugefügtes Material enthalten, dem der Kundenproben ähnelt, die routinemäßig geprüft werden. Zum Beispiel werden reine Materialien, die einer natürlichen Matrix hinzugefügt wurden, häufig viel leichter extrahiert als dieselbe Substanz, die natürlich im Material vorkommt. Falls ein Zweifel besteht, dass dieses geschieht, sollte der Anbieter von Eignungsprüfungen die Eignung der Prüfgegenstände für die Eignungsprüfung bezüglich der zu anzuwendenden Verfahren sicherstellen.

7.3.1.3 Wenn sich durch die Zusammensetzung Prüfgegenstände für die Eignungsprüfung ergeben, bei denen der Zusatzstoff weniger fest gebunden ist als in üblicherweise geprüften Proben oder bei denen er in einer unterschiedlichen Form vorliegt, so kann die Anwendung eines anderen Ansatzes bei der Vorbereitung der Prüfgegenstände für die Eignungsprüfung sinnvoller sein.

7.3.1.4 Die Bestimmung des zugewiesenen Werts anhand der Zusammensetzung ist ein Fall eines allgemeinen Ansatzes für die Charakterisierung zertifizierter Referenzmaterialien nach ISO Guide 35, bei dem ein einzelnes Laboratorium einen zugewiesenen Wert unter Anwendung eines primären Messverfahrens bestimmt. Andere Anwendungen eines primären Verfahrens durch ein einzelnes Laboratorium können angewendet werden, um den zugewiesenen Wert für die Eignungsprüfung zu bestimmen (siehe 7.5).

7.3.2 Wenn der zugewiesene Wert aus der Zusammensetzung des Prüfgegenstands für die Eignungsprüfung berechnet wird, wird die Standardunsicherheit für die Charakterisierung (u_{char}) durch Kombination der Unsicherheiten unter Anwendung eines geeigneten Modells abgeschätzt. Zum Beispiel werden bei Eignungsprüfungen für chemische Messungen die Unsicherheiten normalerweise jene sein, die mit gravimetrischen und volumetrischen Messungen und der Reinheit jeglicher, in der Zusammensetzung verwendeten Materialien verbunden sind. Die Standardunsicherheit des zugewiesenen Werts ($u(x_{pt})$) wird dann nach Gleichung (3) berechnet.

7.4 Zertifiziertes Referenzmaterial

7.4.1 Wenn der Prüfgegenstand für die Eignungsprüfung ein zertifiziertes Referenzmaterial (ZRM) ist, wird sein zertifizierter Eigenschaftswert x_{CRM} als der zugewiesene Wert x_{pt} verwendet.

Die Einschränkungen dieses Ansatzes bestehen darin, dass:

- es aufwändig sein kann, jedem Teilnehmer eine Einheit des zertifizierten Referenzmaterials bereitzustellen;
- ZRM häufig sehr stark verarbeitet werden, um ihre Langzeitstabilität sicherzustellen, wodurch die Austauschbarkeit der Prüfgegenstände für die Eignungsprüfung beeinträchtigt werden kann;
- ein ZRM kann den Teilnehmern bekannt sein, wodurch wichtig wird, dass die Identität des Prüfgegenstands für die Eignungsprüfung verborgen bleibt.

7.4.2 Wenn ein zertifiziertes Referenzmaterial als Prüfgegenstand für die Eignungsprüfung verwendet wird, wird die Standardunsicherheit des zugewiesenen Werts aus den Angaben zur Unsicherheit des Eigenschaftswerts hergeleitet, die dem Zertifikat zu entnehmen sind. Die Angaben aus dem Zertifikat sollten die Komponenten in Gleichung (3) einschließen und ihre Zweckbestimmung sollte für den Zweck des Eignungsprüfungsprogramms geeignet sein.

7.5 Ergebnisse aus einem einzelnen Laboratorium

7.5.1 Ein zugewiesener Wert kann durch ein einzelnes Laboratorium unter Anwendung eines Referenzverfahrens, z. B. eines primären Verfahrens, bestimmt werden. Das angewendete Referenzverfahren sollte vollständig beschrieben und verstanden werden, mit einer vollständigen Angabe der Unsicherheit und dokumentierter metrologischer Rückführbarkeit, die für das Eignungsprüfungsprogramm geeignet sind. Das Referenzverfahren sollte bezüglich aller von Teilnehmern angewendeten Messverfahren austauschbar sein.

7.5.1.1 Der zugewiesene Wert sollte der Mittelwert einer Untersuchung nach Plan sein, die mehr als einen Prüfgegenstand für die Eignungsprüfung oder mehr als eine Messbedingung sowie eine ausreichende Anzahl an Wiederholmessungen verwendet.

7.5.1.2 Die Unsicherheit der Charakterisierung ergibt sich aus dem geeigneten Schätzwert der Unsicherheit für das Referenzverfahren und die Bedingungen der geplanten Untersuchung.

7.5.2 Der zugewiesene Wert x_{pt} des Prüfgegenstands für die Eignungsprüfung kann durch ein einzelnes Laboratorium unter Anwendung eines geeigneten Messverfahrens abgeleitet werden, aus einer Kalibrierung gegen die Referenzwerte eines gut übereinstimmenden zertifizierten Referenzmaterials. Dieser Ansatz geht davon aus, dass das ZRM bezüglich aller von Teilnehmern angewendeten Messverfahren austauschbar ist.

7.5.2.1 Diese Bestimmung erfordert eine Reihe von Prüfungen, die in einem einzelnen Laboratorium, an Prüfgegenständen für die Eignungsprüfung und dem ZRM, unter Anwendung desselben Messverfahrens und unter Wiederholbedingungen durchzuführen sind. Wenn:

x_{CRM} der zugewiesene Wert für das ZRM ist;

x_{pt} der zugewiesene Wert für den Prüfgegenstand für die Eignungsprüfung ist;

d_i die Differenz zwischen den gemittelten Ergebnissen für den Prüfgegenstand für die Eignungsprüfung und für das ZRM bei den Proben mit der Zählnummer i ist;

$\bar{d}$ der Mittelwert der Differenzen d_i ist;

Dann gilt:

$$x_{pt} = x_{CRM} + \bar{d} \quad (4)$$

ANMERKUNG x_{CRM} und $\bar{d}$ sind unabhängig, mit Ausnahme des seltenen Falls, dass das Expertenlaboratorium auch das ZRM hergestellt hat.

7.5.2.2 Die Standardunsicherheit der Charakterisierung wird aus der für die Wertzuweisung verwendeten Messunsicherheit hergeleitet. Dieser Ansatz ermöglicht die Festlegung des zugewiesenen Werts in einer Weise, die auf den zertifizierten Wert des ZRM metrologisch rückführbar ist, mit einer Standardunsicherheit die nach Gleichung (5) berechnet werden kann.

$$u_{char} = \sqrt{u_{CRM}^2 + u_d^2} \quad (5)$$

Das Beispiel in Anhang E.5 veranschaulicht, wie die erforderliche Unsicherheit in dem einfachen Fall berechnet werden darf, in dem der zugewiesene Wert eines Prüfgegenstands für die Eignungsprüfung durch direkten Vergleich mit einem einzelnen ZRM festgelegt wird.

7.5.3 Wenn ein Referenzwert vor Beginn einer Runde eines sequenziellen Eignungsprüfungsprogramms zugewiesen und anschließend unter Anwendung derselben Messeinrichtung überprüft wird, muss die Abweichung zwischen den Werten weniger als das Doppelte der Unsicherheit dieser Abweichung betragen (d. h. die Ergebnisse müssen metrologisch kompatibel sein). In derartigen Fällen darf der Anbieter von Eignungsprüfungen entscheiden, einen Mittelwert der Messungen als zugewiesenen Wert mit der geeigneten Unsicherheit zu verwenden. Falls die Ergebnisse nicht metrologisch kompatibel sind, sollte der Anbieter von Eignungsprüfungen die Ursache der Abweichung untersuchen und geeignete Schritte unternehmen, einschließlich der Anwendung alternativer Verfahren zur Bestimmung des zugewiesenen Werts und seiner Unsicherheit oder des Abbruchs der Runde.

7.6 Konsenswert von Expertenlaboratorien

7.6.1 Zugewiesene Werte können durch einen Ringversuch mit Expertenlaboratorien bestimmt werden, wie in ISO Guide 35 für die Anwendung von Ringversuchen zur Charakterisierung eines ZRM beschrieben. Zuerst werden die Prüfgegenstände für die Eignungsprüfung vorbereitet und für die Verteilung an die Teilnehmer bereit gemacht. Danach werden einige dieser Prüfgegenstände für die Eignungsprüfung zufällig ausgewählt und durch eine Expertengruppe nach einem Protokoll analysiert, das die Anzahl der Prüfgegenstände und der Wiederholmessungen und jegliche anderen maßgebenden Bedingungen festlegt. Von jedem Expertenlaboratorium wird die Bereitstellung einer Standardunsicherheit mit ihren Ergebnissen gefordert.

7.6.2 Wenn die Expertenlaboratorien ein Einzelergebnis berichten und das Messprotokoll von ihnen nicht fordert, ausreichende Angaben zur Messunsicherheit mit den Ergebnissen bereitzustellen oder wenn es in berichteten Ergebnissen oder anderenorts Belege dafür gibt, dass die berichteten Unsicherheiten nicht ausreichend zuverlässig sind, sollte der Konsenswert üblicherweise mit den Verfahren nach 7.7 berechnet werden, die auf den Satz von Ergebnissen der Expertenlaboratorien angewendet werden. Wenn die Expertenlaboratorien mehr als jeweils ein Einzelergebnis berichten (z. B. einschließlich Wiederholmessungen) muss der Anbieter des Eignungsprüfungsprogramms ein alternatives Verfahren zur Bestimmung des zugewiesenen Werts und der zugehörigen Unsicherheit festlegen, das statistisch gültig ist (siehe 4.1.1) und das die Möglichkeit von Ausreißern oder anderen Abweichungen von der erwarteten Verteilung der Ergebnisse berücksichtigt.

7.6.3 Wenn die Expertenlaboratorien Unsicherheiten mit den Ergebnissen berichten, ist die Schätzung eines Werts durch Konsens der Ergebnisse ein komplexes Problem, und es wurde eine breite Vielfalt von Ansätzen vorgeschlagen, einschließlich z. B. gewichteter Mittelwerte, ungewichteter Mittelwerte, Verfahren, die Zuschläge für Überdispersion berücksichtigen sowie Verfahren, die mögliche Ausreißer- oder abweichende Ergebnisse und Unsicherheitsschätzwerte in Betracht ziehen [16]. Der Anbieter von Eignungsprüfungen muss entsprechend ein Schätzverfahren festlegen, das:

- a) Überprüfungen auf Gültigkeit berichteter Unsicherheitsschätzwerte einschließen sollte, z. B. durch Überprüfen, ob berichtete Unsicherheiten der Grund für die beobachtete Ergebnisstreuung sind;
- b) ein Gewichtungsverfahren anwenden sollte, das für das Ausmaß und die Zuverlässigkeit der berichteten Unsicherheiten geeignet ist, das Gleichgewichtung einschließen darf, falls die berichteten Unsicherheiten entweder ähnlich oder von minderer oder unbekannter Zuverlässigkeit sind (siehe 7.6.2);
- c) die Möglichkeit zulassen sollte, dass berichtete Unsicherheiten nicht umfassend der beobachteten Streuung („Überdispersion“) Rechnung tragen, z. B. durch Einschluss eines zusätzlichen Terms um die Überdispersion zu berücksichtigen;
- d) die Möglichkeit unerwarteter Ausreißerwerte für das berichtete Ergebnis oder die Unsicherheit zulassen sollte;
- e) eine fundierte theoretische Grundlage haben sollte;
- f) die für die Zwecke des Eignungsprüfungsprogramms ausreichende Leistung nachgewiesen (z. B. anhand Prüfdaten oder in Simulationen) haben muss.

7.7 Konsenswert aus Teilnehmerergebnissen

7.7.1 Bei diesem Ansatz ist der zugewiesene Wert x_{pt} für den bei einer Runde eines Eignungsprüfungsprogramms verwendeten Prüfgegenstand für die Eignungsprüfung der Lageschätzwert (z. B. robuster Mittelwert, Median oder arithmetischer Mittelwert), der aus den Ergebnissen gebildet wird, die von Teilnehmern in der Runde berichtet wurden, berechnet nach Anhang C mit einem geeigneten Verfahren in Übereinstimmung mit der Planung. Die in 6.2 bis 6.6 beschriebenen Verfahren sollten angewendet werden, um zu bestätigen, dass ausreichende Übereinstimmung besteht, bevor die Ergebnisse zusammengeführt werden.

7.7.1.1 In einigen Fällen kann es der Wunsch des Anbieters von Eignungsprüfungen sein, eine Teilmenge der Teilnehmer einzusetzen, die anhand einiger vorher festgelegter Kriterien, wie Akkreditierungsstatus oder auf Grundlage früherer Leistung als zuverlässig bestimmt wurden. Die Verfahren in diesem Abschnitt gelten für derartige Fälle, einschließlich Erwägungen zur Gruppengröße.

7.7.1.2 Anstatt der Verfahren nach Anhang C dürfen andere Berechnungsverfahren angewendet werden, vorausgesetzt, dass sie eine fundierte statistische Grundlage besitzen und der Bericht das verwendete Verfahren angibt.

7.7.1.3 Die Vorteile dieses Ansatzes sind, dass:

- a) keine zusätzlichen Messungen erforderlich sind, um den zugewiesenen Wert zu erhalten;
- b) der Ansatz bei einer standardisierten Konventionsmessgröße besonders nützlich sein kann, da es häufig kein zuverlässigeres Verfahren zum Erzielen gleichwertiger Ergebnisse gibt.

7.7.1.4 Die Einschränkungen dieses Ansatzes bestehen darin, dass:

- a) die Übereinstimmung zwischen den Teilnehmern ungenügend sein kann;
- b) der Konsenswert aufgrund von allgemein üblichen Verfahrensfehlern einen unbekannten Bias enthalten kann, und dieser Bias sich nicht in der Standardunsicherheit des zugewiesenen Werts widerspiegelt;
- c) der Konsenswert aufgrund der Auswirkung des Bias in Verfahren, die zur Bestimmung des zugewiesenen Werts verwendet werden, verzerrt sein könnte.
- d) es schwierig sein kann, die metrologische Rückführbarkeit des Konsenswerts zu bestimmen. Während das Ergebnis immer auf die Ergebnisse der einzelnen Laboratorien rückführbar ist, kann darüber hinaus eine eindeutige Aussage zur Rückführbarkeit nur getroffen werden, wenn der Anbieter von Eignungsprüfungen über vollständige Angaben über die verwendeten Kalibrierstandards und über die Kontrolle anderer relevanter Verfahrensbedingungen sämtlicher Teilnehmer verfügt, die zum Konsenswert beitragen.

7.7.2 Die Standardunsicherheit des zugewiesenen Werts wird von dem angewendeten Verfahren abhängen. Falls ein vollständiger allgemeiner Ansatz benötigt wird, sollte der Anbieter von Eignungsprüfungen erwägen, Resampling-Verfahren (Bootstrap-Verfahren) anzuwenden, um eine Standardabweichung für den zugewiesenen Wert zu schätzen. Die Literaturhinweise [17],[18] enthalten Einzelheiten zu Bootstrap-Verfahren.

ANMERKUNG Ein Beispiel für ein Bootstrap-Verfahren ist in Anhang E.6 angegeben.

7.7.3 Wenn der zugewiesene Wert als robuster Mittelwert abgeleitet wird, berechnet unter Anwendung der Verfahren nach Anhang C.2, C.3 oder C.5, darf die Standardunsicherheit des zugewiesenen Werts x_{pt} wie folgt abgeschätzt werden:

$$u(x_{pt}) = 1,25 \times \frac{s^*}{\sqrt{p}} \quad (6)$$

Dabei ist s^* die robuste Standardabweichung der Ergebnisse. (Hierbei gilt als „Ergebnis“ für einen Teilnehmer der Mittelwert seiner gesamten Messungen am Prüfgegenstand für die Eignungsprüfung.)

ANMERKUNG 1 In diesem Modell, bei dem der zugewiesene Wert und die robuste Standardabweichung aus Teilnehmerergebnissen bestimmt werden, kann für die Unsicherheit des zugewiesenen Werts angenommen werden, dass die Auswirkungen der Unsicherheit aufgrund von Inhomogenität, Transport und Instabilität eingeschlossen sind.

ANMERKUNG 2 Der Faktor 1,25 beruht auf der Standardabweichung des Medians oder auf der Effizienz des Medians als einem Schätzwert des Mittelwerts in einer großen Reihe von Ergebnissen, die aus einer Normalverteilung genommen wurden. Zwar kann die Effizienz komplexerer robuster Verfahren viel höher sein als die des Medians, was einen niedrigeren Korrekturfaktor als 1,25 rechtfertigt. Dieser Faktor wurde jedoch empfohlen, weil Ergebnisse von Eignungsprüfungen üblicherweise nicht streng normalverteilt sind und unbekannte Anteile von Ergebnissen verschiedener Verteilungen enthalten („kontaminierte Ergebnisse“). Der Faktor 1,25 wird als ein konservativer (hoher) Schätzwert angesehen, um möglicher Kontamination Rechnung zu tragen. Anbieter von Eignungsprüfungen können in der Lage sein, einen kleineren Faktor oder eine andere Gleichung zu begründen, in Abhängigkeit der Erfahrung und des angewendeten robusten Verfahrens.

ANMERKUNG 3 Ein Beispiel für die Verwendung eines zugewiesenen Werts aus Teilnehmerergebnissen wird in Anhang E.3 angegeben.

7.8 Vergleich des zugewiesenen Werts mit einem unabhängigen Referenzwert

7.8.1 Werden die in 7.7 beschriebenen Verfahren zur Bestimmung des zugewiesenen Werts (x_{pt}) angewendet und steht ein zuverlässiger unabhängiger Schätzwert (bezeichnet als x_{ref}) zur Verfügung, z. B. durch die Kenntnis der Herstellung oder durch einen Referenzwert, sollte der Konsenswert x_{pt} mit x_{ref} verglichen werden.

Werden die in 7.3 bis 7.6 beschriebenen Verfahren zur Bestimmung des zugewiesenen Werts angewendet, sollte der aus den Ergebnissen der Runde abgeleitete robuste Mittelwert x^* mit dem zugewiesenen Wert nach jeder Runde eines Eignungsprüfungsprogramms verglichen werden.

Die Differenz wird berechnet als $x_{diff} = (x_{ref} - x_{pt})$ (oder $(x^* - x_{pt})$) und die Standardunsicherheit der Differenz wird wie folgt abgeschätzt:

$$u_{diff} = \sqrt{u^2(x_{ref}) + u^2(x_{pt})} \quad (7)$$

Dabei ist

$u(x_{ref})$ die Unsicherheit des Referenzwerts für den Vergleich; und

$u(x_{pt})$ die Unsicherheit des zugewiesenen Werts.

ANMERKUNG Ein Beispiel für einen Vergleich eines Referenzwerts mit einem Konsenswert ist in Anhang E.7 enthalten.

7.8.2 Beträgt die Differenz mehr als das Doppelte ihrer Standardunsicherheit sollte der Grund dafür untersucht werden. Mögliche Gründe sind:

- Bias im Referenzmessverfahren;
- ein gemeinsames Bias in den Ergebnissen der Teilnehmer;
- die Nichtbeachtung der Einschränkungen des Messverfahrens bei Anwendung des Verfahrens zur Probenherstellung nach 7.3;
- Bias in den Ergebnissen der „Experten“, wenn die Ansätze nach 7.5 oder 7.6 verwendet werden; und
- der Vergleichswert und der zugewiesene Wert sind nicht auf dieselbe metrologische Referenz rückführbar.

7.8.3 Abhängig vom Grund für die Differenz sollte der Anbieter von Eignungsprüfungen entscheiden, ob die Ergebnisse bewertet werden oder nicht, sowie ob (bei kontinuierlichen Eignungsprüfungsprogrammen) die Planung für nachfolgende Eignungsprüfungsprogramme zu ändern ist. Wenn die Differenz ausreichend groß ist, um die Leistungsbeurteilung zu beeinträchtigen oder auf einen erheblichen Bias in den von den Teilnehmern angewendeten Messverfahren hinzudeuten, sollte die Differenz im Bericht für die Runde angegeben werden. In derartigen Fällen sollte die Differenz in der Planung zukünftiger Eignungsprüfungsprogramme berücksichtigt werden.

8 Bestimmung von Kriterien für die Leistungsbewertung

8.1 Ansätze für die Bestimmung von Bewertungskriterien

8.1.1 Der grundlegende Ansatz für alle Zwecke besteht darin, ein Ergebnis (x_i) für einen Prüfgegenstand für die Eignungsprüfung mit einem zugewiesenen Wert (x_{pt}) zu vergleichen. Bei der Bewertung wird die Differenz mit einer Toleranzgrenze für die Messabweichung verglichen. Dieser Vergleich wird für gewöhnlich mit einer standardisierten Leistungskenngröße vorgenommen (z. B. z , z' , ζ , E_n), wie nach 9.4 bis 9.7. Das kann auch erreicht werden, indem die Differenz mit einem festgelegten Kriterium (D oder $D\%$ verglichen mit δ_E) verglichen wird, wie nach 9.3. Ein alternativer Bewertungsansatz ist der Vergleich der Differenz mit der vom Teilnehmer angegebenen Unsicherheit seines Ergebnisses in Verbindung mit der Unsicherheit des zugewiesenen Werts (E_n und ζ).

8.1.2 Falls eine behördliche Anforderung oder ein Gebrauchstauglichkeitsziel als Standardabweichung angegeben ist, darf diese direkt als σ_{pt} verwendet werden. Falls die Anforderung oder das Ziel für eine maximal zulässige Messabweichung gilt, darf dieses Kriterium durch die Eingriffsgrenze dividiert werden, um σ_{pt} zu erhalten. Eine vorgegebene maximal zulässige Abweichung darf direkt als δ_E verwendet werden, für die Verwendung mit D oder $D\%$. Die Vorteile dieses Ansatzes für kontinuierliche Schemata sind:

- a) Leistungs-Scores weisen eine konsistente Interpretation auf, im Sinne der Gebrauchstauglichkeit von einer Runde zur nächsten;
- b) Leistungs-Scores unterliegen nicht der erwarteten Variabilität der Abschätzung der Streuung aus berichteten Ergebnissen.

BEISPIEL Falls ein behördliches Kriterium als maximal zulässige Abweichung festgelegt ist und 3,0 eine Eingriffsgrenze bei der Bewertung mit einem z -Score ist, dann wird das festgelegte Kriterium durch 3,0 geteilt, um σ_{pt} zu bestimmen.

8.1.3 Wenn das Leistungsbewertungskriterium auf Konsenskenngößen aus der aktuellen Runde oder früheren Runden des Eignungsprüfungsprogramms beruht, dann gilt ein robuster Schätzwert der Standardabweichung der Teilnehmerergebnisse als die bevorzugte Kenngröße. Wenn dieser Ansatz verwendet wird, ist es üblicherweise am günstigsten, einen Leistungswert wie den z -Score zu verwenden und die Standardabweichung für die Eignungsbeurteilung (σ_{pt}) auf den berechneten Schätzwert der Standardabweichung zu setzen.

8.2 Werte aus Erkenntnissen von Experten

8.2.1 Die maximal zulässige Abweichung oder die Standardabweichung für die Eignungsbeurteilung darf auf einen Wert festgesetzt werden, der dem Leistungsniveau entspricht, das eine regelsetzende Behörde, eine Akkreditierungsstelle oder die technischen Experten des Anbieters von Eignungsprüfungen als für die Teilnehmer angemessen ansehen.

8.2.2 Eine festgelegte maximal zulässige Abweichung kann in eine Standardabweichung für die Eignungsbeurteilung umgewandelt werden, indem der Grenzwert durch die Anzahl der Vielfachen von σ_{pt} dividiert wird, die verwendet werden, um ein Eingriffssignal (oder ein unzulässiges Ergebnis) festzulegen. Ein festgelegter Wert σ_{pt} kann gleichermaßen in δ_E umgewandelt werden.

8.3 Werte aus Erfahrungen aus früheren Runden eines Eignungsprüfungsprogramms

8.3.1 Die Standardabweichung für die Eignungsbeurteilung (σ_{pt}) und die maximal zulässige Abweichung (δ_E) können anhand der Erfahrung aus früheren Runden der Eignungsprüfung für dieselbe Messgröße mit vergleichbaren Eigenschaftswerten und sofern die Teilnehmer kompatible Messverfahren anwenden, bestimmt werden. Das ist ein nützlicher Ansatz, wenn es keine Übereinstimmung unter den Experten bezüglich der Gebrauchstauglichkeit gibt. Die Vorteile dieses Ansatzes sind die Folgenden:

- Bewertungen werden auf begründbaren Leistungserwartungen beruhen;
- die Bewertungskriterien werden nicht von Runde zu Runde des Eignungsprüfungsprogramms wegen zufälliger Variationen oder Änderungen in der Grundgesamtheit der Teilnehmer variieren;
- die Bewertungskriterien werden nicht zwischen verschiedenen Anbietern von Eignungsprüfungen variieren, wenn zwei oder mehr Anbieter von Eignungsprüfungen für einen Prüf- oder Kalibrierbereich zugelassen sind.

8.3.2 Die Überprüfung von früheren Runden eines Eignungsprüfungsprogramms sollte Überlegungen zur Leistung umfassen, die durch kompetente Teilnehmer erreichbar ist und die nicht durch neue Teilnehmer oder durch Zufallsstreuung aufgrund z. B. kleinerer Gruppengrößen oder anderer Faktoren, die speziell für eine bestimmte Runde gelten, beeinträchtigt wird. Bestimmungen können subjektiv durch Untersuchung der früheren Runden auf Konsistenz vorgenommen werden oder objektiv mit Mittelwerten oder mit einem Regressionsmodell, das auf den Wert der Messgröße anpasst. Die Regressionsfunktion könnte eine Gerade oder eine Kurve [31] sein. Standardabweichungen und relative Standardabweichungen sollten berücksichtigt werden, wobei ihre Auswahl darauf beruht, welche davon über die geeignete Spannweite der Messgrößenniveaus konsistenter ist. Auf diese Weise kann auch eine geeignete maximal zulässige Abweichung erhalten werden.

8.3.3 Wenn das Kriterium für die Leistungsbewertung auf Konsenskenngößen aus früheren Runden eines Eignungsprüfungsprogramms beruht, sollten robuste Schätzwerte der Standardabweichung verwendet werden.

ANMERKUNG 1 Der Algorithmus S (Anhang C.4) bietet eine robuste kombinierte Standardabweichung, die gilt, wenn alle früheren Runden eines betrachteten Eignungsprüfungsprogramms dieselbe erwartete Standardabweichung oder (falls bei der Beurteilung relative Abweichungen verwendet werden) dieselbe relative Standardabweichung aufweisen.

ANMERKUNG 2 Ein Beispiel für das Herleiten eines Werts aus der Erfahrung früherer Runden eines Eignungsprüfungsprogramms ist in Anhang E.8 enthalten.

8.4 Wert aus einem allgemeinen Modell

8.4.1 Der Wert der Standardabweichung für die Eignungsbeurteilung kann aus einem allgemeinen Modell für die Vergleichpräzision des Messverfahrens abgeleitet werden. Dieses Verfahren hat den Vorteil der messgrößenübergreifenden Objektivität und Konsistenz sowie einer empirischen Grundlage. In Abhängigkeit des verwendeten Modells könnte dieser Ansatz als ein Sonderfall für ein Gebrauchstauglichkeitskriterium betrachtet werden.

8.4.2 Jede erwartete Standardabweichung, die für ein allgemeines Modell gewählt wird, muss begründbar sein. Falls sehr großen oder sehr kleinen Anteilen von Teilnehmern Eingriffs- oder Warnsignale zugewiesen werden, sollte der Anbieter von Eignungsprüfungen sicherstellen, dass dies mit dem Zweck des Eignungsprüfungsprogramms in Einklang steht.

8.4.3 Eine spezifische Schätzung, die die Besonderheiten des Messproblems berücksichtigt, ist im Allgemeinen einem allgemeinen Ansatz vorzuziehen. Daraus folgt, dass vor Verwendung eines allgemeinen Modells die Möglichkeit der Verwendung der Ansätze nach 8.2, 8.3 und 8.5 untersucht werden sollte.

BEISPIEL Die Horwitz-Kurve.

Ein übliches allgemeines Modell für chemische Anwendungen wurde von Horwitz [22] beschrieben und durch Thompson [31] modifiziert. Dieser Ansatz stellt ein allgemeines Modell für die Vergleichpräzision von Analyseverfahren dar, das verwendet werden darf, um den folgenden Ausdruck für die Vergleichstandardabweichung herzuleiten:

$$\sigma_R = \begin{cases} 0,22c & \text{wenn } c < 1,2 \times 10^{-7} \\ 0,22c^{0,8495} & \text{wenn } 1,2 \times 10^{-7} \leq c \leq 0,138 \\ 0,01c^{0,5} & \text{wenn } c > 0,138 \end{cases} \quad (8)$$

Dabei ist c der Massenanteil des zu bestimmenden chemischen Agens, wobei $0 \leq c \leq 1$.

ANMERKUNG 1 Das Horwitz-Modell ist empirisch und beruht auf Beobachtungen aus Ringversuchen an vielen Parametern über eine ausgedehnte Zeitspanne. Die σ_R -Werte sind die erwarteten oberen Grenzwerte der Variabilität zwischen Laboratorien, wenn im Ringversuch keine signifikanten Probleme auftraten. Die σ_R -Werte könnten daher als Kriterien für die Bestimmung der Kompetenz in einem Eignungsprüfungsprogramm ungeeignet sein.

ANMERKUNG 2 Ein Beispiel für das Herleiten eines Werts auf Grundlage des modifizierten Horwitz-Modells ist in Anhang E.9 enthalten.

8.5 Werte aus Wiederhol- und Vergleichstandardabweichungen eines früheren Ringversuchs zur Präzision eines Messverfahrens

8.5.1 Wenn das im Eignungsprüfungsprogramm anzuwendende Messverfahren standardisiert ist und Angaben zur Wiederholpräzision (σ_r) und der Vergleichpräzision (σ_R) des Verfahrens verfügbar sind, kann die Standardabweichung für die Eignungsbeurteilung (σ_{pt}) unter Verwendung dieser Angaben wie folgt berechnet werden:

$$\sigma_{pt} = \sqrt{\sigma_R^2 - \sigma_r^2(1 - 1/m)} \quad (9)$$

Dabei ist m die Anzahl der Wiederholmessungen, die jeder Teilnehmer während einer Runde des Eignungsprüfungsprogramms durchzuführen hat.

ANMERKUNG Diese Gleichung wird aus einem grundlegenden Modell der Zufallswirkungen nach ISO 5725-2 abgeleitet.

8.5.2 Wenn Wiederhol- und Vergleichstandardabweichungen vom Mittelwert der Prüfergebnisse abhängen, so sollten die funktionalen Zusammenhänge nach den in ISO 5725-2 beschriebenen Verfahren abgeleitet werden. Diese Zusammenhänge sollten dann verwendet werden, um Werte der Wiederhol- und Vergleichstandardabweichungen zu berechnen, die für den zugewiesenen Wert geeignet sind, der für das Eignungsprüfungsprogramm zu verwenden ist.

8.5.3 Damit die vorstehend genannten Verfahren gültig sind, muss der Ringversuch in Übereinstimmung mit den Anforderungen nach ISO 5725-2 oder nach einem gleichwertigen Verfahren durchgeführt worden sein.

ANMERKUNG Ein Beispiel ist in Anhang E.10 angegeben.

8.6 Werte aus Daten, die in derselben Runde eines Eignungsprüfungsprogramms erzielt wurden

8.6.1 Bei diesem Ansatz wird die Standardabweichung für die Eignungsbeurteilung (σ_{pt}) aus den Ergebnissen der Teilnehmer derselben Runde des Eignungsprüfungsprogramms berechnet. Wenn dieser Ansatz verwendet wird, ist es üblicherweise am günstigsten, einen Leistungswert wie den z-Score zu verwenden. Ein robuster Schätzwert der Standardabweichung der von allen Teilnehmern berichteten Ergebnisse, berechnet mit einem Verfahren nach Anhang C, sollte für gewöhnlich zur Berechnung von σ_{pt} verwendet werden. Im Allgemeinen sind eine Bewertung mit D oder $D\%$ und die Verwendung von δ_E in diesen Fällen nicht geeignet, dennoch kann P_A weiterhin als standardisierter Wert für den messgrößenübergreifenden Vergleich (9.3.6) verwendet werden.

8.6.2 Die Verwendung der Teilnehmerergebnisse kann zu ungeeigneten Leistungsbewertungskriterien führen. Der Anbieter von Eignungsprüfungen sollte sicherstellen, dass der für die Leistungsbewertungen verwendete Wert σ_{pt} gebrauchstauglich ist.

8.6.2.1 Für den Fall, dass die robuste Standardabweichung sehr niedrig ist, sollte der Anbieter von Eignungsprüfungen eine Grenze für den niedrigsten Wert von σ_{pt} ansetzen, der verwendet werden wird. Dieser Grenzwert sollte so gewählt werden, dass der Leistungs-Score in dem Fall, dass die Messabweichung den anspruchsvollsten Verwendungszweck erfüllt, $z < 3,0$ betragen wird.

BEISPIEL In einem Eignungsprüfungsprogramm für ein Gewebe ist eine Messgröße die Anzahl der Fäden je Zentimeter. Die robuste Standardabweichung kann in einigen Runden gering sein (< 1 Fäden je Zentimeter), und Abweichungen unter 4 Fäden/cm werden als unbedeutend betrachtet. Der Anbieter von Eignungsprüfungen bestimmt, dass die robuste Standardabweichung als σ_{pt} verwendet wird, es sei denn, sie beträgt weniger als 1,3 Fäden/cm; in diesem Fall wird $\sigma_{pt} = 1,3$ verwendet.

8.6.2.2 Im Fall, dass die robuste Standardabweichung sehr hoch ist, sollte der Anbieter von Eignungsprüfungen eine Grenze für den höchsten Wert von σ_{pt} ansetzen, der verwendet werden wird, oder eine Grenze für die Messergebnisse, die als „zufriedenstellend“ (kein Signal) bewertet werden können. Dieser Grenzwert sollte so gewählt werden, dass die Ergebnisse, die für den Zweck ungeeignet sind, ein Eingriffssignal erhalten werden.

8.6.2.3 In einigen Fällen darf der Anbieter von Eignungsprüfungen obere oder untere Grenzen für das Intervall der Ergebnisse ansetzen, die als „zufriedenstellend“ (kein Warn- oder Eingriffssignal) bewertet werden können, falls symmetrische Intervalle Ergebnisse einschließen, die nicht gebrauchstauglich sein würden.

BEISPIEL Bei einem behördlichen Eignungsprüfungsprogramm für Wasser, das keine Trinkwassereigenschaft besitzt, legen die Vorschriften fest, dass die Ergebnisse innerhalb $3\sigma_{pt}$ des robusten Mittelwertes der Teilnehmerergebnisse liegen müssen. Da jedoch in einigen Fällen die Spannweite der zulässigen Ergebnisse $0 \mu\text{g/l}$ einschließen könnte, muss jedes Ergebnis, das unter 10 % eines bekannten Werts aus der Herstellung liegt, ein Eingriffssignal (oder „unzulässig“) erzeugen. Ein Prüfgegenstand für die Eignungsprüfung wird mit $4,0 \mu\text{g/l}$ eines regulierten Stoffs hergestellt. Der robuste Mittelwert der Teilnehmer beträgt $3,2 \mu\text{g/l}$ und σ_{pt} beträgt $1,1 \mu\text{g/l}$. Daher ist es für einen Teilnehmer möglich, ein Ergebnis von $0,0 \mu\text{g/l}$ zu übermitteln und innerhalb von $3\sigma_{pt}$ zu bleiben, dennoch wird jedes Ergebnis unter $0,4 \mu\text{g/l}$ als „unzulässig“ bewertet werden.

8.6.3 Die Hauptvorteile dieses Ansatzes sind Einfachheit und gängige Akzeptanz aufgrund erfolgreicher Verwendung in vielen Fällen. Das kann der einzig durchführbare Ansatz sein.

8.6.4 Dieser Ansatz hat mehrere Nachteile:

- a) der Wert von σ_{pt} kann von Runde zu Runde eines Eignungsprüfungsprogramms erheblich schwanken; dadurch wird einem Teilnehmer die Verwendung von Werten des z-Scores zur Suche nach Trends, die über mehrere Runden fortbestehen, erschwert;
- b) Standardabweichungen können unzuverlässig sein, wenn die Anzahl der Teilnehmer im Eignungsprüfungsprogramm klein ist oder wenn Ergebnisse aus verschiedenen Verfahren kombiniert werden. Zum Beispiel kann, für $p = 20$, die Standardabweichung für normalverteilte Daten um bis zu $\pm 30\%$ von ihrem wahren Wert von einer Runde eines Eignungsprüfungsprogramms zur nächsten schwanken;
- c) die Verwendung von Streuungsmaßen, die aus den Daten stammen, kann zu einem annähernd konstanten Anteil von anscheinend zulässigen Scores führen. Generell mangelhafte Leistung wird dann nicht bei der Überprüfung der Scores erkannt, und generell gute Leistung wird dazu führen, dass gute Teilnehmer mangelhafte Scores erhalten;
- d) er liefert keine nützliche Interpretation im Sinne der Eignung für jegliche Endverwendung der Ergebnisse.

ANMERKUNG Beispiele zur Verwendung von Teilnehmerdaten werden im umfangreichen Beispiel in Anhang E.3 bereitgestellt.

8.7 Überwachung der Übereinstimmung zwischen Laboratorien

8.7.1 Als eine Prüfung der Leistung der Teilnehmer und um den Vorteil des Eignungsprüfungsprogramms für die Teilnehmer zu beurteilen, sollte der Anbieter von Eignungsprüfungen ein Verfahren zur Überwachung der Übereinstimmung zwischen Laboratorien, zur Änderungsverfolgung bei der Leistung und zur Sicherstellung der Plausibilität von statistischen Verfahren anwenden.

8.7.2 Die in jeder Runde eines Eignungsprüfungsprogramms erzielten Ergebnisse sollten zur Berechnung von Schätzwerten der Vergleichstandardabweichungen des Messverfahrens (und Wiederholpräzision, falls verfügbar) verwendet werden, unter Anwendung der robusten Verfahren nach Anhang C. Diese Schätzwerte sollten sequentiell oder als Zeitreihen grafisch aufgetragen werden, zusammen mit den Werten der Wiederhol- und Vergleichstandardabweichungen, die in Präzisionsversuchen nach ISO 5725-2 (sofern verfügbar) und/oder aus σ_{pt} erzielt wurden, falls Verfahren nach 8.2 bis 8.4 angewendet werden.

8.7.3 Diese Graphen sollten dann durch den Anbieter von Eignungsprüfungen untersucht werden. Falls die Graphen ergeben, dass die in einer bestimmten Runde der Eignungsprüfung erzielten Präzisionswerte um einen Faktor von zwei oder mehr größer als die erwarteten Werte aus früheren Daten oder Erfahrung sind, sollte der Anbieter von Eignungsprüfungen untersuchen, warum die Übereinstimmung in dieser Runde schlechter als in vorherigen war. Ebenso sollte ein Trend zu besseren oder schlechteren Präzisionswerten eine Untersuchung auf die wahrscheinlichsten Ursachen auslösen.

9 Berechnung von Leistungskenngrößen

9.1 Allgemeine Erwägungen für die Leistungsbestimmung

9.1.1 Kenngrößen für die Leistungsbestimmung müssen mit dem/den Ziel(en) des Eignungsprüfungsprogramms übereinstimmen.

ANMERKUNG Leistungskenngrößen sind am nützlichsten, wenn die Kenngrößen und ihre Ableitung von den Teilnehmern und anderen interessierten Parteien verstanden werden.

9.1.2 Leistungswerte sollten einfach über die Niveaus der Messgrößen und die verschiedenen Runden eines Eignungsprüfungsprogramms übergreifend überprüft werden können.

9.1.3 Die Übereinstimmung der Teilnehmerergebnisse mit den in der Planung des Eignungsprüfungsprogramms verwendeten Annahmen sollte überprüft und bestimmt werden, um aussagekräftige Leistungskenngrößen zu ermöglichen. Das heißt zum Beispiel, dass es keinen Hinweis auf eine Verschlechterung des Prüfgegenstands für die Eignungsprüfung oder auf eine Mischung von Teilnehmer-Grundgesamtheiten oder auf schwerwiegende Verstöße gegen statistische Annahmen zur Beschaffenheit der Daten gibt.

9.1.4 Im Allgemeinen ist es nicht angemessen, Bewertungsverfahren anzuwenden, die absichtlich für einen festen Anteil der Ergebnisse ein Eingriffssignal erzeugen.

9.2 Begrenzung der Unsicherheit des zugewiesenen Werts

9.2.1 Wenn die Standardunsicherheit $u(x_{pt})$ des zugewiesenen Werts im Vergleich zum Leistungsbewertungskriterium groß ist, dann besteht ein Risiko, dass einige Teilnehmer Eingriffs- oder Warnsignale erhalten, jedoch nicht wegen einer Ursache beim Teilnehmer, sondern auf Grund von Ungenauigkeiten bei der Bestimmung des zugewiesenen Werts. Daher muss die Standardunsicherheit des zugewiesenen Werts bestimmt und den Teilnehmern mitgeteilt werden (siehe ISO/IEC 17043:2010, 4.4.5 und 4.8.2).

Falls das folgende Kriterium erfüllt wird, darf die Unsicherheit des zugewiesenen Werts als vernachlässigbar betrachtet werden und braucht nicht in der Interpretation der Ergebnisse der Eignungsprüfungsrunde enthalten zu sein.

$$u(x_{pt}) < 0,3\sigma_{pt} \text{ oder } u(x_{pt}) < 0,1\delta_E \quad (10)$$

ANMERKUNG $0,3\sigma_{pt}$ entspricht $0,1\delta_E$ wenn $|z| \geq 3,0$ ein Eingriffssignal erzeugt.

9.2.2 Wird diesem Kriterium nicht entsprochen, sollte der Anbieter von Eignungsprüfungen folgendes berücksichtigen, wobei sichergestellt wird, dass alle ergriffenen Maßnahmen in Übereinstimmung mit der vereinbarten Politik zur Leistungsbeurteilung für das Eignungsprüfungsprogramm bleiben:

- a) es wird ein Verfahren zur Bestimmung des zugewiesenen Werts gewählt, dessen Unsicherheit dem Kriterium nach Gleichung (10) entspricht;
- b) bei der Auswertung der Ergebnisse des Eignungsprüfungsprogramms wird die Unsicherheit des zugewiesenen Werts berücksichtigt (siehe 9.5 zum z' -Wert oder 9.6 zu ζ -Werten oder 9.7 zu E_n -Werten);
- c) falls der zugewiesene Wert aus Teilnehmerergebnissen abgeleitet wird und die große Unsicherheit von den Differenzen zwischen identifizierbaren Teilgesamtheiten der Teilnehmer herrührt, werden die Werte und Unsicherheiten für jede Teilgesamtheit (z. B. Teilnehmer die unterschiedliche Messverfahren anwenden) gesondert berichtet;

ANMERKUNG Das IUPAC Harmonized Protocol [32] beschreibt ein spezielles Verfahren zur Erkennung von Bimodalität, das auf einer Überprüfung des Kerndichtediagramms mit einer festgelegten Bandbreite beruht.

- d) die Teilnehmer werden informiert, dass die Unsicherheit des zugewiesenen Werts nicht vernachlässigbar ist und Bewertungen beeinträchtigt sein könnten.

Falls keiner der Listeneinträge a) bis d) möglich ist, müssen die Teilnehmer darüber informiert werden, dass kein zuverlässiger zugewiesener Wert bestimmt werden kann und dass keine Leistungswerte bereitgestellt werden können.

ANMERKUNG Die in diesem Abschnitt angegebenen Verfahren, werden in den Anhängen E.3 und E.4 dargelegt.

9.3 Schätzwerte der Abweichung (Messabweichung)

9.3.1 Sei x_i das Ergebnis (oder der Mittelwert der Wiederholmessungen), das von einem Teilnehmer i für die Messung einer Eigenschaft des Prüfgegenstands für die Eignungsprüfung bei einer Runde eines Eignungsprüfungsprogramms berichtet wurde. Dann kann ein einfaches Maß der Leistung des Teilnehmers als Differenz zwischen dem Ergebnis x_i und dem zugewiesenen Wert x_{pt} berechnet werden.

$$D_i = x_i - x_{pt} \quad (11)$$

D_i kann als die Messabweichung für dieses Ergebnis interpretiert werden, bis zu dem Ausmaß, in dem der zugewiesene Wert als ein konventioneller oder Referenzgrößenwert betrachtet werden kann.

Die Differenz D_i kann in denselben Einheiten wie der zugewiesene Wert angegeben werden oder als eine prozentuale Differenz, berechnet als:

$$D_i\% = 100(x_i - x_{pt})/x_{pt}\% \quad (12)$$

9.3.2 Die Differenz D oder $D\%$ wird üblicherweise mit einem Kriterium δ_E auf Grundlage der Gebrauchstauglichkeit oder aus der Erfahrung früherer Runden eines Eignungsprüfungsprogramms verglichen; das Kriterium wird hier als δ_E , eine Toleranzgrenze für die Messabweichung, angegeben. Wenn $-\delta_E < D < \delta_E$, dann wird die Leistung als „zufriedenstellend“ (bzw. „kein Signal“) betrachtet. (Dasselbe Kriterium gilt für $D\%$, abhängig von der Angabe von δ_E .)

9.3.3 δ_E ist eng verknüpft mit dem für z-Scores verwendeten σ_{pt} (siehe 9.4), wenn σ_{pt} anhand der Gebrauchstauglichkeit oder anhand von Erwartungen aus früheren Runden bestimmt wird. Der Zusammenhang wird durch das Bewertungskriterium für z-Scores bestimmt. Falls z. B. $z \geq 3$ ein Eingriffssignal erzeugt, dann gilt $\delta_E = 3\sigma_{pt}$ oder entsprechend $\sigma_{pt} = \delta_E/3$. Verschiedene Angaben von δ_E sind bei Eignungsprüfungen für medizinische Anwendungen und bei Leistungsspezifikationen für Messverfahren und Produkte gebräuchlich.

9.3.4 Der Vorteil von D als eine Leistungskenngröße und δ_E als ein Leistungskriterium besteht darin, dass Teilnehmer ein intuitives Verständnis dieser Kenngrößen besitzen, weil diese direkt mit der Messabweichung verknüpft und als Kriterium zur Bestimmung der Gebrauchstauglichkeit gebräuchlich sind. Der Vorteil von $D\%$ besteht darin, dass es intuitiv verständlich ist, für das Niveau der Messgröße standardisiert ist und dass es im Zusammenhang mit üblichen Gründen für Abweichungen steht (z. B. fehlerhafte Kalibrierung oder Bias bei Verdünnung).

9.3.5 Die Nachteile sind, dass das in vielen Ländern oder Gebieten der Messtechnik bei Eignungsprüfungen nicht gebräuchlich ist; sowie dass D nicht standardisiert ist, um ein einfaches Durchsuchen von Berichten nach Eingriffssignalen in Eignungsprüfungsprogrammen mit mehreren Analyten zu ermöglichen oder wenn die Gebrauchstauglichkeitskriterien entsprechend des Niveaus der Messgröße variieren können.

ANMERKUNG Die Verwendung von D und $D\%$ geht im Allgemeinen von einer symmetrischen Verteilung der Teilnehmerergebnisse aus, in dem Sinne, dass die zulässige Spannweite $-\delta_E < D < \delta_E$ beträgt.

9.3.6 Zu Zwecken des Vergleichs zwischen Messgrößenniveaus, bei denen die Gebrauchstauglichkeitskriterien variieren können; oder für die runden- oder messgrößenübergreifende Kombination, können D und $D\%$ in einen standardisierten Leistungswert transformiert werden, der die Differenzen im Verhältnis zu den Leistungskriterien für die Messgröße angibt. Dafür wird für jedes Ergebnis der „prozentuale Anteil der zulässigen Abweichung“ (P_A) wie folgt berechnet:

$$P_{Ai} = (D_i / \delta_E) \times 100 \% \quad (13)$$

Daher weist $P_A \geq 100 \%$ oder $P_A \leq -100 \%$ auf ein Eingriffssignal (oder „unzulässige Leistung“) hin.

ANMERKUNG 1 P_A -Werte können über Niveaus und verschiedene Runden eines Eignungsprüfungsprogramms übergreifend verglichen oder in Diagrammen rückverfolgt werden. Diese Leistungswerte ähneln in Verwendung und Interpretation z-Scores mit einem gebräuchlichen Bewertungskriterium, wie $z \leq -3$ oder $z \geq 3$ für Eingriffssignale.

ANMERKUNG 2 Variationen dieser Kenngrößen werden allgemein verwendet, insbesondere bei medizinischen Anwendungen, bei denen eine höhere Häufigkeit der Eignungsprüfungen und eine große Anzahl von Analyten üblich sind.

ANMERKUNG 3 Es kann geeignet sein, den absoluten Wert von P_A zu verwenden, um durchgehend zulässige (oder unzulässige) Ergebnisse im Verhältnis zum zugewiesenen Wert widerzuspiegeln.

9.4 z-Scores

9.4.1 Der z-Score für ein Ergebnis einer Eignungsprüfung, x_i , wird berechnet als:

$$z_i = \frac{(x_i - x_{pt})}{\sigma_{pt}} \quad (14)$$

Dabei ist

- x_{pt} der zugewiesene Wert; und
- σ_{pt} die Standardabweichung für die Eignungsbeurteilung.

9.4.2 Die konventionelle Interpretation von z-Scores ist wie folgt (siehe ISO/IEC 17043:2010, B.4.1.1):

- ein Ergebnis, das $|z| \leq 2,0$ ergibt, wird als zufriedenstellend angesehen;
- ein Ergebnis, das $2,0 < |z| < 3,0$ ergibt, wird als ein Warnsignal auslösend angesehen;
- ein Ergebnis, das $|z| \geq 3,0$ ergibt, wird als nicht zufriedenstellend (oder Eingriffssignal) angesehen.

Teilnehmern sollte geraten werden, ihre Messverfahren nach Warnsignalen zu überprüfen, falls sie auf ein entstehendes oder wiederholtes Problem hinweisen.

ANMERKUNG 1 Bei einigen Anwendungen verwenden Anbieter von Eignungsprüfungen den Wert 2,0 als ein Eingriffssignal für z-Werte.

ANMERKUNG 2 Das Kriterium σ_{pt} wird üblicherweise so gewählt, dass die vorstehende Interpretation ermöglicht wird, die häufig für Eignungsbeurteilungen verwendet wird und ebenfalls gebräuchlichen Regelkarten-Grenzen sehr ähnlich ist.

ANMERKUNG 3 Die Begründung für die Verwendung von Grenzen von 2,0 und 3,0 für z-Scores lautet wie folgt: Von korrekt durchgeführten Messungen wird angenommen, dass sie Ergebnisse erzeugen, die durch eine Normalverteilung mit dem Mittelwert x_{pt} und der Standardabweichung σ_{pt} beschrieben werden können (nach der Transformation, falls nötig). Dann werden z-Scores normalverteilt sein, mit einem Mittelwert von null und einer Standardabweichung von 1,0. Unter diesen Umständen würde erwartet werden, dass nur ungefähr 0,3 % der Scores außerhalb der Spannweite $-3,0 \leq z \leq 3,0$ liegen und nur ungefähr 5 % außerhalb der Spannweite $-2,0 \leq z \leq 2,0$. Weil die Wahrscheinlichkeit dafür, dass z außerhalb von $\pm 3,0$ liegt so gering ist, ist es unwahrscheinlich, dass Eingriffssignale zufällig auftreten werden, wenn keine wirklichen Probleme bestehen, so dass es wahrscheinlich ist, dass es eine identifizierbare Ursache einer Anomalie gibt, wenn ein Eingriffssignal ausgelöst wird.

ANMERKUNG 4 Die zugrundeliegende Annahme dieser Interpretation gilt nur für eine angenommene Verteilung von kompetenten Laboratorien und nicht für jegliche Annahmen zur Verteilung der beobachteten Ergebnisse. Es braucht keine Annahme zu den beobachteten Ergebnissen an sich getroffen zu werden.

ANMERKUNG 5 Falls die wahre Variabilität zwischen Laboratorien kleiner als σ_{pt} ist, werden die Wahrscheinlichkeiten für Fehlklassifizierung verringert.

ANMERKUNG 6 Wenn die Standardabweichung für die Eignungsbeurteilung durch die Verfahren entweder nach 8.2 oder nach 8.4 festgelegt ist, kann sie beträchtlich von der (robusten) Standardabweichung der Ergebnisse abweichen, und die Anteile der Ergebnisse, die außerhalb $\pm 2,0$ und $\pm 3,0$ liegen kann erheblich von 5 % bzw. 0,3 % abweichen.

9.4.3 Der Anbieter von Eignungsprüfungen muss geeignete Rundungsverfahren für die berichteten z-Scores bestimmen, auf Grundlage der Anzahl signifikanter Stellen für das Ergebnis und für den zugewiesenen Wert und die Standardabweichung für die Eignungsprüfung. Die Rundungsregeln müssen in den Angaben enthalten sein, die den Teilnehmern zur Verfügung stehen.

ANMERKUNG Bei z-Scores sind mehr als zwei Nachkommastellen selten nützlich.

9.4.4 Wenn die Standardabweichung der Teilnehmerergebnisse als σ_{pt} verwendet wird und Eignungsprüfungsprogramme große Anzahlen von Teilnehmern umfassen, kann der Anbieter von Eignungsprüfungen den Wunsch haben, die Normalität der Verteilung unter Verwendung von tatsächlichen Ergebnissen oder z-Scores zu überprüfen. Im anderen Extremfall, wenn es nur eine kleine Anzahl von Teilnehmern gibt, kann es sein, dass kein Eingriffssignal ausgelöst wird. In diesem Fall können grafische Verfahren, die Leistungswerte über mehrere Runden kombinieren, nützlichere Angaben zur Leistung der Teilnehmer bieten als die Ergebnisse von einzelnen Runden.

9.5 z'-Scores

9.5.1 Wenn Bedenken bezüglich der Unsicherheit eines zugewiesenen Werts, $u(x_{pt})$, bestehen, z. B. wenn $u(x_{pt}) > 0,3\sigma_{pt}$, dann kann die Unsicherheit durch Erweiterung des Nenners des Leistungswerts berücksichtigt werden. Diese Kenngröße wird als z'-Score bezeichnet und wird wie folgt (bei Verwendung der Notation nach 9.4) berechnet:

$$z'_i = \frac{x_i - x_{pt}}{\sqrt{\sigma_{pt}^2 + u^2(x_{pt})}} \quad (15)$$

ANMERKUNG Wenn x_{pt} und/oder σ_{pt} aus Teilnehmerergebnissen berechnet werden, ist der Leistungswert mit einzelnen Teilnehmerergebnissen korreliert, weil Einzelergebnisse eine Auswirkung auf sowohl einen robusten Mittelwert als auch die robuste Standardabweichung haben. Die Korrelation für einen einzelnen Teilnehmer ist abhängig von der Wichtung, die diesem Teilnehmer in der kombinierten Kenngröße zukommt. Aus diesem Grund stellen Leistungswerte, die die Unsicherheit des zugewiesenen Werts enthalten ohne einen Zuschlag für die Korrelation einzuschließen, Unterschätzungen der Scores dar, die erzielt würden, wenn die Kovarianz enthalten wäre. Wenn z. B. $u(x_{pt}) = 0,3\sigma_{pt}$, dann gibt es eine Unterschätzung von ungefähr 10 % des z'-Werts. Daher kann Gleichung (15) verwendet werden, wenn x_{pt} und/oder σ_{pt} aus Teilnehmerergebnissen bestimmt werden.

9.5.2 *D*- und *D*%-Scores können ebenfalls modifiziert werden, um die Unsicherheit des zugewiesenen Werts mit der folgenden Gleichung zu berücksichtigen, um δ_E auf δ_E' zu erweitern:

$$\delta_E' = \sqrt{\delta_E^2 + U^2(x_{pt})} \quad (16)$$

Dabei ist $U(x_{pt})$ die erweiterte Unsicherheit des zugewiesenen Werts x_{pt} , die mit einem Erweiterungsfaktor $k = 2$ berechnet wird.

9.5.3 z' -Scores können auf dieselbe Weise interpretiert werden, wie z -Scores (siehe 9.4); dabei werden dieselben kritischen Werte von 2,0 und 3,0 verwendet, abhängig von der Planung für das Eignungsprüfungsprogramm. In gleicher Weise würden dann die *D*- und *D*Scores mit δ_E' (siehe 9.3) verglichen werden.

9.5.4 Ein Vergleich der Gleichungen für den z -Score und den z' -Score in 9.4 beziehungsweise 9.5 zeigt, dass die z' -Scores für eine Runde eines Eignungsprüfungsprogramms stets kleiner sein werden als die entsprechenden z -Werte; zwischen beiden Werten liegt ein konstanter Faktor von

$$\frac{\sigma_{pt}}{\sqrt{\sigma_{pt}^2 + u^2(x_{pt})}}$$

Wird die Leitlinie für die Ermittlung der Grenzen der Unsicherheit des zugewiesenen Werts in 9.2.1 erfüllt, wird dieser Faktor innerhalb folgender Grenzen liegen:

$$0,96 < \frac{\sigma_{pt}}{\sqrt{\sigma_{pt}^2 + u^2(x_{pt})}} < 1,00$$

Daher werden die z' -Scores in diesem Fall nahezu mit den z -Scores übereinstimmen und es darf die Schlussfolgerung gezogen werden, dass die Unsicherheit des zugewiesenen Werts bei der Leistungsbewertung vernachlässigbar ist.

Wird die Leitlinie in 9.2.1 für die Unsicherheit des zugewiesenen Werts nicht erfüllt, können die z' -Scores und die z -Scores größenordnungsmäßig so weit voneinander abweichen, dass einige der z -Scores die kritischen Werte von 2,0 oder 3,0 überschreiten und daher „Warnsignale“ oder „Eingriffssignale“ ergeben, während die entsprechenden z' -Scores diese kritischen Werte nicht überschreiten und daher auch keine Signale erzeugen.

Im Allgemeinen darf in Fällen, in denen der zugewiesene Wert und/oder σ_{pt} nicht aus Teilnehmerergebnissen bestimmt wird, z' bevorzugt werden, weil bei Erfüllung des Kriteriums nach 9.2.1 die Differenz zwischen z und z' vernachlässigbar sein wird.

9.6 Zeta-Scores (ζ)

9.6.1 Zeta-Scores können nützlich sein, wenn ein Ziel für das Eignungsprüfungsprogramm darin besteht, die Fähigkeit eines Teilnehmers zu bewerten, Ergebnisse zu erzielen, die näher am zugewiesenen Wert liegen als ihre angegebene Unsicherheit.

In der Notation nach 9.4 werden ζ -Scores wie folgt berechnet:

$$\zeta_i = \frac{x_i - x_{pt}}{\sqrt{u^2(x_i) + u^2(x_{pt})}} \quad (17)$$

Dabei ist

$u(x_i)$ der teilnehmereigene Schätzwert der Standardunsicherheit des eigenen Ergebnisses x_i ; und

$u(x_{pt})$ die Standardunsicherheit des zugewiesenen Werts x_{pt} .

ANMERKUNG 1 Wenn der zugewiesene Wert, x_{pt} , als der Konsenswert aus Teilnehmerergebnissen berechnet wird, korreliert x_{pt} mit den einzelnen Teilnehmerergebnissen. Die Korrelation für einen einzelnen Teilnehmer ist abhängig von der Wichtung, die diesem Teilnehmer beim zugewiesenen Wert zukommt und in geringerem Umfang bei der Unsicherheit des zugewiesenen Werts. Aus diesem Grund stellen Leistungswerte, die die Unsicherheit des zugewiesenen Werts enthalten ohne einen Zuschlag für die Korrelation einzuschließen, Unterschätzungen der Werte dar, die erzielt würden, wenn die Kovarianz enthalten wäre. Das Unterschätzen ist nicht schwerwiegend, sofern die Unsicherheit des zugewiesenen Werts gering ist; wenn robuste Verfahren angewendet werden, ist es am wenigsten schwerwiegend für die am weitesten außen liegenden Teilnehmer, die am wahrscheinlichsten nachteilige Leistungswerte erhalten. Gleichung (17) kann daher mit Konsenskenngrößen ohne Korrelationskorrektur verwendet werden.

ANMERKUNG 2 ζ -Scores unterscheiden sich von E_n -Scores (9.7) durch die Verwendung von Standardunsicherheiten $u(x_i)$ und $u(x_{pt})$ an Stelle von erweiterten Unsicherheiten $U(x_i)$ und $U(x_{pt})$. ζ -Scores über 2 oder unter -2 können durch systematisch abweichende Verfahren hervorgerufen werden oder durch mangelhafte Schätzung der Messunsicherheit durch den Teilnehmer. ζ -Scores liefern daher eine strenge Beurteilung des vom Teilnehmer mitgeteilten, vollständigen Ergebnisses.

9.6.2 Die Verwendung von ζ -Scores ermöglicht die direkte Beurteilung, ob Laboratorien fähig sind, korrekte Ergebnisse zu liefern, d.h. Ergebnisse, die mit x_{pt} innerhalb ihrer Messunsicherheiten übereinstimmen. ζ -Scores können unter Verwendung derselben kritischen Werte von 2,0 und 3,0 wie bei z-Scores oder mit den Vielfachen des bei der Schätzung der erweiterten Unsicherheit verwendeten Erweiterungsfaktors des Teilnehmers interpretiert werden. Ein nachteiliger ζ -Score kann jedoch auf eine große Abweichung für x_i von x_{pt} hinweisen, eine teilnehmerseitige Unterschätzung der Unsicherheit oder eine Kombination von beidem.

ANMERKUNG Es kann nützlich sein, wenn der Anbieter von Eignungsprüfungen zusätzliche Angaben über die Gültigkeit berichteter Unsicherheiten bereitstellt. Nützliche Leitlinien für derartige Beurteilungen werden in 9.8 vorgeschlagen.

9.6.3 ζ -Scores können als Hilfestellung bei der Verbesserung der Leistung von Teilnehmern wie folgt zusammen mit z-Scores verwendet werden. Falls ein Teilnehmer z-Scores erzielt, die wiederholt den kritischen Wert von 3,0 überschreiten, kann es ratsam sein, ihr Prüfverfahren schrittweise zu untersuchen und eine Unsicherheitsbewertung für dieses Verfahren zu erhalten. Die Unsicherheitsbewertung wird die Schritte in dem Verfahren, bei dem die größten Unsicherheiten auftreten, so identifizieren, dass der Teilnehmer sehen kann, wo er seine Anstrengungen erhöhen muss, um eine Verbesserung zu erreichen. Falls die ζ -Scores des Teilnehmers ebenfalls den kritischen Wert von 3,0 wiederholt überschreiten, deutet das darauf hin, dass die Unsicherheitsbewertung des Teilnehmers nicht sämtliche signifikanten Unsicherheitsquellen umfasst (d. h., dass sie einen wichtigen Faktor übersehen). Wenn im umgekehrten Fall ein Teilnehmer wiederholt z-Scores ≥ 3 , aber ζ -Scores < 2 erzielt, zeigt das, dass der Teilnehmer die Unsicherheit seiner Ergebnisse genau beurteilt haben kann, dass aber seine Ergebnisse die für das Eignungsprüfungsprogramm erwartete Leistung nicht erfüllen. Das kann z.B. der Fall sein bei einem Teilnehmer, der als Messverfahren ein Screening-Verfahren anwendet, während die anderen Teilnehmer quantitative Verfahren anwenden. Falls der Teilnehmer die Unsicherheit seiner Ergebnisse als ausreichend erachtet, ist keine Maßnahme erforderlich.

ANMERKUNG Wenn ein ζ -Score einzeln verwendet wird, kann er nur als eine Prüfung dessen interpretiert werden, ob die Unsicherheit des Teilnehmer mit der besonderen beobachteten Abweichung übereinstimmt und kann nicht als Hinweis für die Gebrauchstauglichkeit der Ergebnisse eines besonderen Teilnehmers interpretiert werden. Die Bestimmung der Gebrauchstauglichkeit könnte gesondert erfolgen (z.B. durch den Teilnehmer oder eine Akkreditierungsstelle), durch Untersuchung der Abweichung $(x - x_{pt})$ oder der kombinierten Standardunsicherheiten im Vergleich mit einer Zielunsicherheit.

9.7 E_n -Scores

9.7.1 E_n -Scores können nützlich sein, wenn ein Ziel für das Eignungsprüfungsprogramm darin besteht, die Fähigkeit eines Teilnehmers zu bewerten, Ergebnisse zu erzielen, die nah am zugewiesenen Wert, innerhalb ihrer angegebenen erweiterten Unsicherheit liegen. Diese Kenngröße ist bei Eignungsprüfungen für Kalibrierungen gebräuchlich, sie kann jedoch für andere Arten der Eignungsprüfung verwendet werden.

Die Leistungskenngröße wird wie folgt berechnet:

$$(E_n)_i = \frac{x_i - x_{pt}}{\sqrt{U^2(x_i) + U^2(x_{pt})}} \quad (18)$$

Dabei ist

x_{pt} der in einem Referenzlaboratorium bestimmte, zugewiesene Wert;

$U(x_{pt})$ die erweiterte Unsicherheit des zugewiesenen Werts x_{pt} ;

$U(x_i)$ die erweiterte Unsicherheit eines Teilnehmerergebnisses x_i .

ANMERKUNG Die direkte Kombination von erweiterten Unsicherheiten entspricht nicht der Anforderung nach ISO/IEC Guide 98-3 und entspricht nicht der Berechnung einer kombinierten erweiterten Unsicherheit, sofern nicht die Erweiterungsfaktoren und die effektiven Freiheitsgrade für $U(x_i)$ und $U(x_{pt})$ identisch sind.

9.7.2 E_n -Scores sollten mit Vorsicht interpretiert werden, weil sie Verhältnisse zweier gesonderter (aber zusammenhängender) Leistungsmaße sind. Der Zähler ist die Abweichung des Ergebnisses vom zugewiesenen Wert und seine Interpretation wird in 9.3 dargelegt. Der Nenner ist eine kombinierte erweiterte Unsicherheit, die nicht größer als die Abweichung im Zähler sein sollte, falls der Teilnehmer $U(x_i)$ korrekt bestimmt hat und falls der Anbieter von Eignungsprüfungen $U(x_{pt})$ korrekt bestimmt hat. Daher könnten die Werte von $E_n \geq 1,0$ oder $E_n \leq -1,0$ eine Anforderung für die Überprüfung der Unsicherheitsschätzwerte oder für die Berichtigung eines Messproblems aufzeigen; gleichzeitig sollten $-1,0 < E_n < 1,0$ nur dann als ein Hinweis auf erfolgreiche Leistung dienen, sofern die Unsicherheiten gültig sind und die Abweichung $(x_i - x_{pt})$ niedriger ist, als von den Kunden des Teilnehmers benötigt.

ANMERKUNG Die Interpretation von E_n -Scores kann zwar schwierig sein, das verhindert jedoch nicht ihre Verwendung. Das Einbeziehen von unsicherheitsbezogenen Angaben in die Interpretation von Ergebnissen von Eignungsprüfungen kann eine Hauptrolle bei der Verbesserung des Verständnisses der Teilnehmer hinsichtlich Messunsicherheit und ihrer Bewertung spielen.

9.8 Bewertung der Teilnehmerunsicherheiten bei der Prüfung

9.8.1 Die verstärkte Anwendung von ISO/IEC 17025 führte zu einem besseren Verständnis der Messunsicherheit. Die Verwendung der Unsicherheitsabschätzungen der Teilnehmer bei der Leistungsbewertung war in Eignungsprüfungsprogrammen in verschiedenen Bereichen der Kalibrierung, wie etwa mit E_n -Scores, üblich; es war jedoch bei der Eignungsprüfung von Prüflaboratorien nicht üblich. Die ζ -Scores nach 9.6 und die E_n -Scores nach 9.7 sind Möglichkeiten für die Bewertung von Ergebnissen gegen die angegebene Unsicherheit.

9.8.2 Einige Anbieter von Eignungsprüfungen haben den Nutzen erkannt, Laboratorien dazu aufzufordern, die Unsicherheit von Ergebnissen in der Eignungsprüfung zu berichten. Das kann selbst dann zweckmäßig sein, wenn die Unsicherheiten nicht für die Bewertung durch Scores verwendet werden. Es gibt mehrere Gründe, derartige Informationen zu sammeln:

- a) Akkreditierungsstellen können sicherstellen, dass Teilnehmer Unsicherheiten berichten, die ihrem Anwendungsbereich der Akkreditierung entsprechen;

- b) Teilnehmer können ihre berichtete Unsicherheit gegen die anderer Teilnehmer überprüfen, um die Konsistenz (sofern vorhanden) zu beurteilen und dadurch eine Möglichkeit zu gewinnen, zu erkennen, ob ihre Berechnung der Unsicherheit nicht alle maßgebenden Komponenten erfasst oder einige Komponenten überschätzt;
- c) Eignungsprüfungen können angewendet werden, um Angaben zur Unsicherheit zu bestätigen; das ist am einfachsten, wenn die Unsicherheit mit dem Ergebnis berichtet wird.

ANMERKUNG Ein Beispiel für die Datenanalyse beim Berichten von Unsicherheiten ist in Anhang E.4 enthalten.

9.8.3 Wenn x_{pt} mit den Verfahren nach 7.3 bis 7.6 bestimmt wird und $u(x_{pt})$ das Kriterium nach 9.2.1 erfüllt, ist es unwahrscheinlich, dass ein Teilnehmerergebnis eine geringere Standardunsicherheit als diese aufweisen wird, so dass $u(x_{pt})$ als untere Grenze für das Screening, als u_{min} bezeichnet, verwendet werden könnte. Falls der zugewiesene Wert anhand von Teilnehmerergebnissen bestimmt wird (7.7), sollte der Anbieter von Eignungsprüfungen praktische Screening-Grenzen für u_{min} bestimmen.

ANMERKUNG Falls $u(x_{pt})$ eine Variabilität aufgrund von Inhomogenität oder Instabilität einschließt, könnte $u(x_i)$ des Teilnehmers kleiner als u_{min} sein.

9.8.4 Es ist ebenfalls unwahrscheinlich, dass die vom Teilnehmer berichtete Standardunsicherheit größer ist, als das 1,5-Fache der robusten Standardabweichung der Teilnehmer ($1,5 s^*$), so dass dieser Wert als praktische obere Grenze für das Screening berichteter Unsicherheiten, als u_{max} bezeichnet, verwendet werden könnte.

ANMERKUNG Der Faktor 1,5 ist die obere Grenze der Variabilität in Standardabweichungen, die für eine Konsens-Standardabweichung mit zehn oder mehr Ergebnissen auf Grundlage der Quadratwurzel der Perzentile der F -Verteilung erwartet werden kann. Anbieter von Eignungsprüfungen, die dieses Verfahren übernehmen, können möglicherweise einen unterschiedlichen Multiplikator verwenden wollen.

9.8.5 Falls u_{min} oder u_{max} oder andere Kriterien verwendet werden, um abweichende Unsicherheiten zu identifizieren, sollte der Anbieter von Eignungsprüfungen das den Teilnehmern erklären und verdeutlichen, dass eine berichtete Unsicherheit, $u(x_i)$, auch dann gültig sein kann, wenn sie niedriger ist als u_{min} oder größer ist als u_{max} ; wenn dieser Fall eintritt, sollten die Teilnehmer und alle interessierten Parteien das Ergebnis oder den Unsicherheitsschätzwert überprüfen. Gleichmaßen kann eine berichtete Unsicherheit größer als u_{min} und kleiner als u_{max} sein und dennoch nicht gültig sein. Sie sind ausschließlich informative Indikatoren.

9.8.6 Anbieter von Eignungsprüfungen können auch auf außergewöhnlich hohe oder niedrige Unsicherheiten aufmerksam machen, die z. B. beruhen auf:

- festgelegten Quantilen für die berichteten Unsicherheiten (z. B. unter dem 5. Perzentil und über dem 95. Perzentil der berichteten Standardunsicherheiten oder der erweiterten Unsicherheiten);
- Grenzen, die auf einer angenommenen Verteilung beruhen, mit einer Skala, die auf der Streuung der berichteten Unsicherheiten beruht;
- einer erforderlichen Messunsicherheit.

ANMERKUNG Da es unwahrscheinlich ist, dass Unsicherheiten normalverteilt sind, ist wahrscheinlich eine Transformation nötig, wenn Grenzen verwendet werden, die auf der annähernden oder zugrundeliegenden Normalität beruhen; z. B. weisen Box-Whisker-Plot-Grenzen, die auf dem Interquartilabstand beruhen, nur dann eine probabilistische Interpretation auf, wenn die Verteilung annähernd normal ist.

9.9 Kombinierte Leistungs-Scores

9.9.1 Es ist üblich, dass innerhalb einer Runde eines Eignungsprüfungsprogramms, Ergebnisse an mehr als einem Prüfgegenstand für die Eignungsprüfung oder für mehr als eine Messgröße ermittelt werden. In diesem Fall sollten die Ergebnisse für jeden Prüfgegenstand für die Eignungsprüfung und jede Messgröße entsprechend 9.3 bis 9.7 ausgewertet werden; d. h., dass die Ergebnisse für jeden Prüfgegenstand für die Eignungsprüfung und jede Messgröße gesondert ausgewertet werden sollten.

9.9.2 Es gibt Anwendungen, bei denen zwei oder mehr Prüfgegenstände für die Eignungsprüfung mit speziell zugeschnittenen Werten in ein Eignungsprüfungsprogramm aufgenommen werden, um andere Aspekte der Leistung zu messen, z. B. um die Wiederholpräzision, die systematische Messabweichung oder die Linearität zu untersuchen. Zum Beispiel können zwei ähnliche Prüfgegenstände für die Eignungsprüfung in einem Eignungsprüfungsprogramm verwendet werden, mit der Absicht, sie in einem Youden-Diagramm zu verwenden, wie nach 10.5. In solchen Fällen sollte der Anbieter von Eignungsprüfungen den Teilnehmern vollständige Beschreibungen der statistischen Planung und der statistischen Verfahren übermitteln, die zur Anwendung kommen.

9.9.3 Die grafischen Verfahren nach Abschnitt 10 sollten verwendet werden, wenn Ergebnisse an mehr als einem Prüfgegenstand für die Eignungsprüfung oder für mehrere Messgrößen ermittelt werden, vorausgesetzt, sie stehen in engem Zusammenhang und/oder wurden mit demselben Verfahren erzielt. Diese Verfahren kombinieren Leistungswerte in einer Weise, die hohe Einzelwerte nicht verbirgt, und sie können zusätzliche Informationen über die Leistungsfähigkeit der Teilnehmer aufzeigen, zum Beispiel Korrelationen zwischen Ergebnissen für verschiedene Messgrößen, die aus den Tabellen der einzelnen Werte nicht erkennbar sind.

9.9.4 Bei Eignungsprüfungsprogrammen, die eine große Anzahl von Messgrößen einbeziehen, kann eine Zählung oder ein Anteil der Anzahl von Eingriffs- und Warnsignalen verwendet werden, um die Leistung zu bewerten.

9.9.5 Kombinierte Leistungs-Scores oder Prämierungs- oder Straf-Scores sollten nur mit Vorsicht verwendet werden, weil es schwierig sein kann, die den Werten zugrundeliegenden statistischen Annahmen zu beschreiben. Während kombinierte Leistungswerte für Ergebnisse bei verschiedenen Prüfgegenständen bei derselben Messgröße die erwarteten Verteilungen aufweisen können und bei der Erkennung von durchgängigem Bias nützlich sein können, können hingegen über verschiedene Messgrößen übergreifende, bei demselben oder bei verschiedenen Prüfgegenständen gemittelte oder zusammengefasste Werte den Bias in den Ergebnissen für einzelne Messgrößen verschleiern. Das Verfahren der Berechnung, die Interpretation und die Grenzen sämtlicher verwendeter kombinierter Scores oder Straf-Scores müssen darum den Teilnehmern deutlich dargelegt werden.

10 Grafische Verfahren zur Beschreibung von Leistungswerten

10.1 Anwendung grafischer Verfahren

Der Anbieter von Eignungsprüfungen sollte üblicherweise die sich in jeder Runde des Eignungsprüfungsprogramms ergebenden Leistungswerte zur Erstellung von grafischen Darstellungen, wie die nach 10.2 und 10.3, verwenden. Die Verwendung von Leistungswerten, wie P_A -, z -, z' -, $\bar{z}$ - oder E_n -Scores, in diesen Graphen hat den Vorteil, dass standardisierte Achsen verwendet werden können, wodurch sich Darstellung und Deutung vereinfachen. Die grafischen Darstellungen sollten den Teilnehmern zur Verfügung gestellt werden, so dass jeder Teilnehmer erkennen kann, wie sich seine Ergebnisse in Relation zu denjenigen der anderen Teilnehmer anordnen. Es können Buchstaben- oder Zahlenschlüssel für die Teilnehmer verwendet werden, so dass jeder Teilnehmer seine eigenen Ergebnisse erkennen kann, jedoch nicht in der Lage ist, zuzuordnen, welcher Teilnehmer welches andere Ergebnis erzielte. Die grafischen Darstellungen können auch vom Anbieter von Eignungsprüfungen und von jeder Akkreditierungsstelle verwendet werden, damit diese die Gesamt-Wirksamkeit des Eignungsprüfungsprogramms beurteilen und erkennen können, ob das zur Leistungsbewertung verwendete Kriterium überprüft werden muss.

10.2 Histogramme von Ergebnissen oder Leistungswerten

10.2.1 Das Histogramm ist ein übliches statistisches Instrument und bei zwei verschiedenen Punkten in der Analyse von Eignungsprüfungsergebnissen nützlich. Die grafische Darstellung ist in der vorläufigen Analysephase nützlich, um zu überprüfen, ob die statistischen Annahmen begründbar sind oder ob eine Anomalie vorliegt, z. B. eine bimodale Verteilung, ein großer Anteil von Ausreißern oder unübliche Schiefe, die nicht vorhergesehen wurde.

Histogramme können auch bei Berichten für das Eignungsprüfungsprogramm von Nutzen sein, um die Leistungswerte zu beschreiben oder um Ergebnisse zu vergleichen, z. B. anhand von unterschiedlichen Verfahren oder verschiedenen Prüfgegenständen für die Eignungsprüfung. Histogramme sind insbesondere bei einzelnen Berichten für kleine oder mittelgroße Eignungsprüfungsprogramme hilfreich (weniger als 100 Teilnehmer), weil sie Teilnehmern ermöglichen, ihre Leistung im Vergleich mit anderen Teilnehmern zu beurteilen, z. B. indem ein Block in einem senkrechten Balken hervorgehoben wird, um die Ergebnisse eines Teilnehmers darzustellen oder, bei kleinen Eignungsprüfungsprogrammen (weniger als 50 Teilnehmer), indem für jeden Teilnehmer individualisierte Diagrammzeichen verwendet werden.

10.2.2 Histogramme können mithilfe tatsächlicher Teilnehmerergebnisse oder mit Leistungs-Scores erstellt werden. Teilnehmerergebnisse haben den Vorteil des direkten Zusammenhangs zu den übermittelten Daten und sie können ohne weitere Berechnung oder Transformation vom Leistungswert zur Messabweichung beurteilt werden. Auf Leistungswerten beruhende Histogramme haben den Vorteil, dass sie in direktem Zusammenhang mit Leistungsbewertungen stehen und einfach über Messgrößen und Runden eines Eignungsprüfungsprogramms übergreifend verglichen werden können.

Der Wertebereich und die Klassenbreite, die für ein Histogramm verwendet werden, müssen für jeden Datensatz auf Grundlage der Variabilität und der Anzahl der Ergebnisse bestimmt werden. Es ist häufig möglich, das auf Grundlage von Erfahrung mit Eignungsprüfungen durchzuführen, aber in den meisten Fällen wird es erforderlich sein, die Klassenbildung nach der ersten Einsichtnahme anzupassen. Falls im Histogramm Leistungswerte verwendet werden, ist es hilfreich, wenn eine Skala auf Grundlage der Standardabweichung für die Eignungsbeurteilung vorliegt sowie Schnittpunkte für Warn- und Eingriffssignale.

10.2.3 Die Schrittweiten für Skala und Diagramm sollten so gewählt werden, dass Bimodalitäten erkannt werden können (falls vorhanden) ohne dass, aufgrund der Auflösung der Messergebnisse oder einer kleinen Anzahl von Ergebnissen, falsche Warnsignale ausgelöst werden.

ANMERKUNG 1 Das Erscheinungsbild von Histogrammen reagiert empfindlich auf die gewählte Breite der Klassen und auf die Lage der Klassengrenzen (bei konstanter Breite der Klassen ist dies größtenteils abhängig vom Anfangspunkt). Falls die Breite der Klassen zu gering ist, wird das Diagramm viele kleine Modi aufweisen; zu große und nennenswerte Modi nahe der Hauptmenge können nicht ausreichend unterscheidbar sein. Das Erscheinungsbild schmaler Modi und die relativen Höhen benachbarter Balken können sich beträchtlich ändern, wenn der Anfangspunkt oder die Klassenbreite geändert wird, insbesondere wenn der Datensatz klein ist und/oder eine Clusterbildung aufweist.

ANMERKUNG 2 Ein Beispiel für ein Histogramm-Diagramm ist in Anhang E.3 angegeben.

10.3 Kerndichtediagramme

10.3.1 Ein Kerndichtediagramm bietet eine gleichmäßige Kurve, die die allgemeine Form der Verteilung eines Datensatzes beschreibt. Der dem Kerndichteschätzer zugrunde liegende Gedanke ist, dass jeder Datenpunkt durch eine festgelegte Verteilung ersetzt wird (üblicherweise Normalverteilung), punktzentriert und mit einer Standardabweichung σ_k ; σ_k wird üblicherweise als „Bandbreite“ bezeichnet. Diese Verteilungen werden zusammenaddiert und die sich daraus ergebende Verteilung, auf eine Flächeneinheit skaliert, ergibt einen „Dichteschätzwert“, der als eine gleichmäßige Kurve grafisch dargestellt werden kann.

10.3.2 Die folgenden Schritte können bei der Erstellung eines Kerndichtediagramms befolgt werden. Es wird angenommen, dass ein Datensatz X , bestehend aus p -Werten $x_1, x_2, \dots, x_p$ in das Diagramm aufgetragen wird. Dies sind üblicherweise Teilnehmerergebnisse, aber sie können auch aus den Ergebnissen abgeleitete Leistungswerte sein.

- i) Auswählen einer geeigneten Bandbreite σ_k . Zwei Möglichkeiten sind besonders nützlich:
- a) bei der allgemeinen Überprüfung wird $\sigma_k = 0,9 s^*/p^{0,2}$ festgelegt; dabei ist s^* eine robuste Standardabweichung der Werte $x_1, \dots, x_p$, die mit Verfahren nach Anhang C.2 oder C.3 berechnet werden.
 - b) für die Untersuchung des Datensatzes auf grobe Modi, die verglichen mit dem Leistungsbeurteilungskriterium wichtig sind, wird festgelegt $\sigma_k = 0,75 \sigma_{pt}$, wenn z - oder ζ -Scores verwendet werden, oder es wird festgelegt $\sigma_k = 0,25 \delta_E$, wenn D oder $D\%$ verwendet wird.

ANMERKUNG 1 Die vorstehende Option a) folgt Silverman[30], der s^* auf Grundlage des normierten Interquartilabstands ($nIQR$) empfiehlt. Andere Regeln zur Bandbreitenauswahl, die ähnliche Ergebnisse bieten, umfassen die Regeln von Scott [29], bei denen der Multiplikator von 0,9 durch 1,06 ersetzt wird. [29] beschreibt ein nahezu optimales, aber sehr viel komplexeres Verfahren zur Bandbreitenauswahl. In der Praxis sind die Unterschiede bei der visuellen Überprüfung gering und die Auswahl hängt von der Software-Verfügbarkeit ab.

ANMERKUNG 2 Die vorstehende Option b) folgt dem IUPAC-Leitfaden [32].

- ii) Es wird ein Bereich für die grafische Darstellung festgelegt q_{min} bis q_{max} , so dass $q_{min} \leq \min(x_1, \dots, x_p) - 3\sigma_k$ und $q_{max} \geq \max(x_1, \dots, x_p) + 3\sigma_k$.
- iii) Es wird eine Anzahl von Punkten, n_k , für die gezeichnete Kurve ausgewählt. $n_k = 200$ ist üblicherweise ausreichend, es sei denn, es gibt extreme Ausreißer im Bereich des Diagramms.
- iv) Die Lage des grafischen Auftrags q_1 bis q_{n_k} wird berechnet aus:

$$q_i = q_{min} + (i - 1) \frac{(q_{n_k} - q_1)}{n_k - 1} \quad (19)$$

- v) Die n_k Dichten h_1 bis h_{n_k} werden berechnet aus:

$$h_i = \frac{1}{p} \sum_{j=1}^p \varphi \left(\frac{x_j - q_i}{\sigma_k} \right) \text{ für } i = 1 \text{ bis } i = n_k \quad (20)$$

Dabei ist

$\varphi (\cdot)$ die Bezeichnung der Standard-Normaldichte.

- vi) Es wird h_i gegen q_i aufgetragen.

ANMERKUNG 1 Es kann hilfreich sein, die Lage der einzelnen Datenpunkte dem Diagramm hinzuzufügen. Das wird am häufigsten getan, indem die Lage unter die gezeichnete Dichtekurve als kurze vertikale Marker (im Englischen mitunter als „rug“ bezeichnet) gezeichnet wird, aber es kann auch erreicht werden, indem die Datenpunkte an den geeigneten Punkten entlang der berechneten Dichtekurve aufgetragen werden.

ANMERKUNG 2 Dichtediagramme werden am besten mithilfe von Software erstellt. Die vorstehende schrittweise Berechnung kann in einer Tabellenkalkulation für mittlere Datensatzgrößen erfolgen. Sowohl proprietäre als auch frei erhältliche Software umfassen Dichte-Diagramme auf Grundlage von ähnlich voreingestellter Bandbreitenauswahl. Fortgeschrittene Software-Implementierungen von Dichte-Diagrammen können diesen Algorithmus oder schnellere Berechnungen auf Grundlage von Faltungsverfahren verwenden.

ANMERKUNG 3 Beispiele für Kerndichtediagramme sind in den Anhängen E.3, E.4, und E.6 enthalten.

10.3.3 Die Kurvenform wird als ein Hinweis auf die Verteilung verwendet, anhand der die Daten aufgetragen wurden. Ausgeprägte Modi erscheinen als gesonderte Spitzen. Ausreißerwerte erscheinen als gesonderte Spitzen, die deutlich von der Hauptmenge der Daten getrennt sind.

ANMERKUNG 1 Eine Dichtekurve reagiert empfindlich auf die gewählte Bandbreite σ_k . Ist die Bandbreite zu gering, wird das Diagramm viele kleine Modi zeigen; zu große und nennenswerte Modi nahe der Hauptmenge können möglicherweise nicht ausreichend unterscheidbar sein.

ANMERKUNG 2 Wie Histogramme werden Dichtediagramme am besten mit mittleren bis großen Datensätzen verwendet, weil kleine Datensätze (zehn oder weniger) zufällige geringfügige Ausreißer oder offensichtliche Modi enthalten können, insbesondere wenn eine robuste Standardabweichung der Bandbreite zugrunde gelegt wird.

10.4 Säulendiagramme von standardisierten Leistungswerten

10.4.1 Säulendiagramme sind ein geeignetes Verfahren zur Darstellung der Leistungswerte für eine Anzahl ähnlicher Merkmale auf einem Graphen. Sie zeigen auf, ob gemeinsame Merkmale in den Werten eines Teilnehmers auftreten, z. B. falls ein Teilnehmer mehrere hohe z-Scores erzielt, die auf allgemein mangelhafte Leistung hinweisen, kann dieser Teilnehmer einen positiven Bias aufweisen.

10.4.2 Zur Erstellung eines Säulendiagramms werden die standardisierten Leistungswerte in ein Säulendiagramm aufgetragen, wie nach Bild E.10, worin die Werte für jeden Teilnehmer gruppiert sind. Andere standardisierte Leistungswerte, wie $D\%$ oder P_A können für denselben Zweck grafisch dargestellt werden.

10.4.3 Wenn in einer Runde eines Eignungsprüfungsprogramms Wiederholbestimmungen durchgeführt werden, können die Ergebnisse zur Berechnung eines Graphen von Präzisionsmaßen verwendet werden; z. B. k -Kenngrößen nach ISO 5725-2 oder ein zugehöriges Maß, das gegen die robuste mittlere Standardabweichung, wie die im Algorithmus S (Anhang C.4) festgelegt, skaliert ist.

ANMERKUNG Ein Beispiel für ein Säulendiagramm mit z-Scores ist in Anhang E.11 enthalten.

10.5 Youden-Diagramm

10.5.1 Wurden während einer Runde eines Eignungsprüfungsprogramms zwei ähnliche Prüfgegenstände für die Eignungsprüfung geprüft, bietet das Youden-Diagramm ein sehr aussagekräftiges grafisches Verfahren zur Untersuchung der Ergebnisse. Es kann für den Nachweis der Korrelation (oder der Unabhängigkeit) von Ergebnissen an unterschiedlichen Prüfgegenständen für die Eignungsprüfung und für die Lenkung von Untersuchungen der Gründe von Eingriffssignalen hilfreich sein.

10.5.2 Der Graph wird konstruiert, indem die Teilnehmerergebnisse oder die z-Scores, die an einem der Prüfgegenstände für die Eignungsprüfung ermittelt wurden, gegen die bei dem anderen Prüfgegenstand für die Eignungsprüfung ermittelten Teilnehmerergebnisse oder z-Scores aufgetragen werden. Vertikale und horizontale Linien werden üblicherweise so aufgetragen, dass sie vier Quadranten von Werten ergeben, um die Auswertung zu unterstützen. Die Linien liegen bei den zugewiesenen Werten oder bei den Medianen für die beiden Verteilungen der Ergebnisse oder bei 0, wenn z-Scores aufgetragen werden.

ANMERKUNG Für eine geeignete Interpretation von Youden-Diagrammen ist es wichtig, dass die beiden Prüfgegenstände für die Eignungsprüfung ähnliche (oder identische) Messgrößenniveaus aufweisen, damit die Art jeder systematischen Messabweichung in diesem Bereich des Messintervalls dieselbe ist. Youden-Diagramme können bei sehr unterschiedlichen Niveaus einer Messgröße beim Vorhandensein einer konsistenten systematischen Abweichung nützlich sein; wenn jedoch ein Kalibrierfehler über die gesamte Spannweite von Messgrößenniveaus nicht durchgehend positiv oder negativ ist, können sie irreführend sein.