採点、テスト分析、テスト得点の報告における
品質管理に関する ITC ガイドライン
日本語版

ITC Guidelines on Quality Control in Scoring, Test Analysis, and Reporting of Test Scores Japanese Version.

Translation authorized
by the Japan Association for Research on Testing.

2013 年 10 月 12 日, version 1.2
最終版
文書番号：ITC-G-QC-20131012

日本語版の作成は、荒井清佳（大学入試センター）、劉東岳（学研教育総合研究所）、渡邉誠一（日本人事試験研究センター）によって行われた。邦訳過程の監修は繁桝算男（日本テスト学会会長）が行った。

Secretary-General: Secretary@InTestCom.org.

公式の採択

本ガイドラインは、オランダのアムステルダムで 2012 年 7 月に開催された ITC 理事会において正式に採択された。

オンラインでの公開

本ガイドラインは、アムステルダムで 2012 年 7 月に開催された ITC 総会で正式に公開され、それ以降 ITC のウェブサイト（http://www.intestcom.org）においてオンラインで見ることができるようになっている。

出版物での公開

本ガイドラインは出版物としてはまだ公開されていない。

本文書を引用する際は次のように記載すること：

International Test Commission (2012). International Guidelines on Quality Control in Scoring, Test Analysis, and Reporting of Test Scores. [www.intestcom.org]

謝辞

本ガイドラインは Avi Allalouf が ITC 理事会のために準備した。著者は本プロジェクトにおける Marise Born の有益な支援に感謝するとともに、旧バージョンをレビューし、本ガイドラインの策定に貢献した次の方々に謝意を表する。

Alvaro Arce-Ferrer, Pearson Vue, USA
James Austin, Ohio State University, USA
Jo-Anne Baird, Oxford University, UK
Giulia Balboni, University of Valle d'Aosta, ITALY
Helen Baron, Independent Consultant, UK
Dave Bartram, SHL, UK
Marise Born, Erasmus University Rotterdam, NETHERLANDS
James Butcher, University of Minnesota, USA
Janet Carlson, Buros Center for Testing, USA
Iain Coyne, University of Nottingham, UK
Kurt Geisinger, University of Nebraska-Lincoln, USA
Ron Hambleton, University of Massachusetts, USA
John Hattie, University of Auckland, NEW ZEALAND
Fred Leong, Michigan State University, USA
Jason Lamprinou, European University, CYPRUS
Tom Oakland, University of Florida, USA
Fred Oswald, Rice University, USA
Christopher Rush, Wireless Generation, Inc., USA

多くの品質管理の手続きを開発し、日常的に用いている NITE(National Institute for Testing and Evaluation)、特に Scoring and Equating Department の同僚にも謝意を表する。

概要

以下に示す品質管理（QC: Quality Control）に関するガイドライン（QC ガイドライン）は、テスティングにおける採点、分析、報告（scoring, analysis and reporting; SAR）の効率性、精密さ、正確さを高めることを目的としている。このガイドラインには二通りの利用方法がある。一つは、採点、分析、報告の品質管理のためのガイドラインとして単独で利用する方法である。もう一つは、テストの使用に関する ITC 国際ガイドライン（2000）の部分的な拡張とみなして利用する方法である。

QC ガイドラインは、指定された試験日に複数のテストフォームが用いられるような大規模テストの運用に焦点を当てている。しかしながら、他の多様なテスト場面（例：進路指導や自己啓発のためのテスト）やアセスメント技術（例：多枝選択式テスト、パフォーマンス評価、構造化・非構造化面接、集団活動の評価）にも、また、アセスメントが行われるほとんどの場面においても（例：教育目的や雇用アセスメントセンターにおいて）利用できる。QC ガイドラインの中には、個人あるいは集団に対して実施される標準化テストに関連するものもあれば、より広く応用できるものもある（例：臨床や教育、職業に関するテスティングにおいて）。多くの職業（例：医療やリハビリテーション、法医学、教育、雇用関連）がアセスメント活動を行っており、ここでもまた QC ガイドラインは非常に有用である。QC ガイドラインはどのような実施形式のテストにも適用できる（紙筆テストや、インターネットを介する、または、介さないコンピュータ化されたテストなど）。

目次

はじめに

ねらいと目標

標準化と正確さは、テスティングの全ての段階、すなわち、テストの開発・実施から、採点、分析、得点の解釈、得点の報告の各段階において、最も重要なものである。採点、分析、得点の報告に関わる全ての者には、採用企業・組織、心理学関連の学協会、大学、所管官庁、法人などの利害関係者に対して正当化できるような専門的な規準を守る責任がある。専門家として携わる者は、どの段階でも起こりうる過失（error）を意識するべきであり、過失を防ぎ、対処するために規準に従って行動しなくてはならない。正解番号の間違いによる得点の誤りや、素点から標準得点への変換での誤り、得点の計算ミス、間違った相手への得点報告、得点についての誤った解釈などはいずれも起こるべきではない過失の例である。過ちを犯すのが人間であるが、適切な品質管理の手続きに則ることで過失を最小限にすべきである。また、テストに携わる者は、品質管理の実践について幅広い知識を持つべきである。というのも、それらはテストの正確な使用には不可欠なことであるからである。我々が少しずつ進めている「継続的な品質の改善（CQI: Continuous Quality Improvement)」の分野にも、本ガイドラインが貢献するものと信じている。

以下に示す品質管理（QC: Quality Control）に関するガイドライン（QC ガイドライン）は、テスティングにおける採点、分析、報告（SAR ： scoring, analysis and reporting）の効率性、精密さ、正確さを高めることを目的としている。このガイドラインには二通りの利用方法がある。一つは、採点、分析、報告の品質管理のためのガイドラインとして単独で利用する方法である。もう一つは、テストの使用に関する ITC ガイドライン（2001）の一部分の拡張とみなして利用する方法である。読者は ITC ガイドラインや AERA, APA & NCME によるスタンダード(2011)とともに、関連する国際的な規準や地域の規準に精通していることが推奨される。

品質管理に関するガイドラインが想定している読者

QC ガイドラインは大規模テスト、主に達成度や潜在能力（好みの指標や自己報告による測定とは異なり）を測定するテストを対象とする。そのため、大規模な教育テストや採用試験における大規模な能力測定は特に適用の対象である。しかしながら、多くの内容は、小規模なアセスメントや他の種類のアセスメントにも応用できる。

QC ガイドラインは次の業務に携わる者を対象にしている：

- テストの設計及び開発
- テストの実施
- テストの採点
- 項目分析およびテスト分析（規準化と等化を含む）
- テストのセキュリティの維持
- テストの解釈
- テスト結果の報告及び受検者へのフィードバックの提供
- テスト使用者に対する訓練と監督
- テストデータを扱うためのコンピュータシステム及びプログラムの設計
- 政策立案（法令制定者を含む）
- テストの出版

品質管理についてよく知ることは、テスティングに関わるどの専門家にとっても非常に重要なことである。QC ガイドラインは、テスティングの分野に実際に携わる者による専門的な使用を主目的としているが、グッドプラクティスに関する基本原則を具体化したものであるので、現場あるいは研究室において研究目的だけにテストを使用する者にも意義がある。

文脈上の要因・国際的な要因

QC ガイドラインの目的は、仕事でテストを扱う世界中の専門職の人々が、その地域に特化した品質規準を開発するのに役立つことである。文脈上の要因に関しては、QC ガイドラインを地域レベルで解釈したり、実際に利用したりする際には、その地域の法令や規準、規制、顧客とテスト会社間の契約などを考慮しなければならない。例えば、国によっては受検者の個人情報に関する守秘義務が法律で定められている。

さまざまな過失　品質管理に関するガイドラインの必要性

採点、分析、報告の過程で生じる過失は、測定する領域がどのようなものであっても（心理や教育、職業、態度など）深刻な結果をもたらすことが考えられる。例えば、採点ミスが多ければ、得点の意味や信頼性が影響を受ける――テスト得点の信頼性はほぼ確実に低くなり、予測的妥当性もまた低くなる。過失により、行動が病的な人が行動が正常な人と

誤って判定される場合もあれば、能力のある応募者が仕事を得られなかったり、学校で不適切なクラス分けが行われたりする場合もあるだろう。過失はまた、不適切な教育プログラムを割り当てたり、必要な知識や技能がないにもかかわらず免許や認可を与えたりするなど、誤った教育的介入を引き起こすこともあるだろう。過失は、得点の報告を大幅に遅らせることがあり、その結果として教育機関での手続きが間に合わないなど重大な問題を引き起こすこともありうる。要するに、過失は、深刻な、損害を与えるような結果をもたらしうる。過失はまた、世間の教育テストや心理テストへの信頼を失わせることがあり、メディアに注目されれば、そのテストの信憑性を減らすことになるであろう。また、場合によっては、過失によってテスト機関や教育機関、テスティングの専門家やサイコメトリシャン、さらには人を雇用しようとしている会社等に対して訴訟が起こされることもあるだろう。

テスティングに携わる専門家（心理学者やサイコメトリシャン、進路指導員など）は、所属機関、受検者、テスト会社そしてメディアからの圧力を受けるおそれがある。いずれも、短期間で費用をかけずにテストを開発して、テスト得点は迅速に報告して利用できるようにしろと要求してくる。品質規準を維持するためには、テスティングの過程を短縮させたり、急がせたり、あるいは段階のいくつかを省略させようとする者からの圧力に屈してはいけない。例えば、採点、分析、得点の報告を短期間で行う契約になっているような場合には、厳しい圧力がかかってくる。

テストの開発、採点（特に答案が大量にある場合）、分析、得点の報告のように連続する段階から成り、それぞれが先行する段階に大きく依存しているような長期にわたる過程においては、過失が生じる可能性が大きい。品質規準の使用は過失を防ぐ助けとなる。品質規準は日常的にモニターし、定期的に更新されるべきである。

品質管理の定義

本文書の目的から考えると、品質管理とは、採点からテスト分析、テスト結果の報告までの全ての段階において高い品質規準を維持するのに役立ち、それによって過失を最小限に抑え、測定の信頼性を高めるように設計された体系的なプロセスと定義できる。

他の職業での例

品質管理の手続きは、エンジニアリング、航空会社、ソフトウェア開発、医療など、他の多くの職業でも用いられる。医療の分野に関しては、病院で起きる過失の背景にある要因が参考になる。薬の不適切な保管、治療介入の複雑さ、新しい技術、連携の不備、チームワークの悪さ、明確な安全方針の欠如などがある。このような例はテスティングの分野でも類似するものがあり、実施と評価の過程においてこのような過失が起こりうる。

品質管理に関するガイドラインの構成

QC ガイドラインは次の二つの部から構成される。

1. 一般原則――採点、分析、得点報告に先立って、考慮され合意されるべき一般的事項
2. 段階別のガイドライン

これらの他に、簡単な要約、参考文献がある。

終わりに

本ガイドラインで述べられている提言に加えて、一般的なガイドラインや提案を列挙する。新しいテストが導入されるたびに、全ての過程を段階的に行う詳細なシミュレーションが実際になされるべきである（Texas Education Agency et al., 2004 を見よ）。新しいテスティングの手続きを実行し、評価するのはそれからである。そのようなシミュレーションは品質管理に関する規準が次に改訂される際に何らかの知見を与えるであろう。さらに、採点、分析、得点報告は連続した段階からなるものであり、その各段階は先行する段階が首尾よく行われることを前提としている。それゆえ、QC ガイドラインに基づいた品質管理チェックリストを作る場合、先行する段階が完了することなく次の段階に進めることができないようにするべきである。コンピュータ化は、容易に、意識させることなく、効果的に QC の手続きを標準化し、修正し、コントロールする最も論理的な方法のように見える。しかしながら、コンピュータ化による利点は広く受け入れられているけれども、研究の心得のある熟練した者が品質管理の手続きを考案し、適応させ、評価する必要がある。

ガイドライン

品質管理に関するガイドラインの範囲

QC ガイドラインは、指定された試験日に複数のテストフォームが用いられるような大規模テストの運用に焦点を当てている。しかしながら、他の多様なテスト（例：進路指導のためのテストや自己啓発のためのテスト）やアセスメント技術（例：多枝選択式テスト、パフォーマンス評価、構造化・非構造化面接、集団活動の評価）にも、アセスメントが行われるほとんどの場面においても（例：教育目的や雇用アセスメントセンターにおいて）利用できる。QC ガイドラインの項目の中には、個人あるいは集団に対して実施される標準化されたテストに関連するものもあれば、より広く応用できるものもある（例：臨床や教育、職業に関するテスティングにおいて）。多くの職業（例：医療やリハビリテーション、法医学、教育、雇用関連）がアセスメント活動を行っており、ここでもまた QC ガイドラインは非常に有用である。

QC ガイドラインは、どのような実施形式のテストにも適用できる（紙筆テストや、インターネットを介する、または、介さないコンピュータ化されたテストなど）。テストの開発やテストの選択、テストの実施は QC ガイドラインの範囲ではない。しかしながら、採点、分析、得点報告における QC ガイドラインが有用か、あるいはうまく適用できるかどうかは、テストそのものが適切であるか、得点が信頼できるものであるか、得点が明確な結果を予測できるものであるかどうかによる。品質管理にリソースを割り当てることは、責任のある行動、説明責任、公正性――いずれも倫理規定の重要な要素であるが――への投資である。

第 1 部：一般原則

1.1. 現在使用されている品質管理に関する規準の確認

1.1.1. 所属機関や国において、現在どのような品質管理に関するガイドラインがあるかを調べ、必要であればテストの実施前にそのテストに合わせた品質管理の手続きを策定する。テストの実施方法に変更が加えられるたびに、また定期的な点検として時々、ガイドラインを見直し、更新し、修正する。

1.1.2. テストを実施する前に、適切な品質管理の手続きが整っていることを確認する。

1.1.3. 新しいテストを扱う場合は、採点、分析、報告の過程全体の試行的なシミュ

レーションを実施するようにする。試行的な実施が行われない場合には、最初の実施を予備的な実施として扱い、次の実施までに改良できるよう準備しておく。

1.1.4. そのテストに合わせた規準がない場合には作成する。

1.1.5. 新しいテストを開発する際には、そのテストに合わせた規準を作成する。

1.2. 基本的な準備と関係者間の合意

テストを実施する前に、テストに携わる者、すなわちテストの開発や実施、採点、等化、結果の解釈、妥当性の検証、結果の報告等の担当者の間で、基本方針について意見を一致させておく。各自の責任や役割はさまざまであっても、テストに携わる者―――業者や顧客や共同事業者であっても―――の関わる仕事は連携されているべきである。異なる役割を担う者の間の連携によって、テストの品質が高まり、テストがより適切な目的で使用されるようになるはずである。

1.2.1. テスティングに関わる全ての利害関係者を確認し、テスティングの各過程において意思決定をする責任が誰にあるか合意しておく。

1.2.2. テストの使用目的を定め、明示する（選抜のため、達成度を測定するため、研究のためなど）。

1.2.3. 採点、分析、報告の過程の実施計画について合意しておく。

1.2.4. 例えば関連情報をあるチームから他のチームへ知らせたり、テストについての詳細（テストの構造や正解番号など）を開発チームから分析チームへ伝達したりするのに最適な方法など、個人間あるいはチーム間（複数のチームが関わる場合）での最適な連絡方法を構築する。

1.2.5. テスティングについて顧客と連絡をとる最適な方法を構築する。

1.2.6. アセスメントデータを採点、分析、報告の過程の責任者へ渡す方法を定める。アセスメントデータとは、例えば、紙筆テストでは光学式文字読み取り装置やスキャナーで読み取ったデータ、コンピュータを用いたテストでは電子的に得られたデータのことである。

1.2.7. （下位テストが用いられる場合には）下位テストの配点を明確にし、その理由を示す。データを受け取った後に配点を修正する準備もしておくべきだが、修正は、理論とテストの目的に従ってのみ行う。

1.2.8. 採点指示、つまり、正しい解答に対する得点や誤っている解答をどう扱うかなどについて、合意しておく。データを受け取った後で採点指示を修正する準備もしておく。

1.2.9. 採点尺度を定め、尺度得点の範囲を定める。

1.2.10. 欠損データをどう扱うか決める（例：受検者が問題項目を見落としたり、解答

を記入する際に誤って 1 行抜かしたりした場合や、あるいは評価者が特定の受検者の評価を忘れてしまったり、測定を繰り返せない状況で、標準化されていない方法で評価してしまったりした場合)。

1.2.11. 異なる版のテスト得点を共通尺度上に示す必要がある場合には、等化モデル、デザイン、必要なサンプルサイズを等化方法とともに明確に述べる。

1.2.12. 判定基準を設定する場合には、そのモデルやデザイン、必要なサンプルサイズについて明確に述べる。

1.2.13. 受検者と関係機関にどの得点を報告するべきか、また、得点分布や得点の使用に関してどのような付加的情報を伝えるべきか、報告の詳細さの程度について合意しておく。

1.2.14. データのプライバシーに関して、法的な制約を守りながら、テスト結果をどの個人、団体・機関が受け取るべきかを定める。

1.2.15. 結果報告において、その他の個人的な情報(テストの内容が修正されているか、何問の問題に答えたか、障害に対してどのような対応がなされたか、など)を提供できるか、あるいは提供すべきかどうかを定める。

1.2.16. テスティングの過程全体について、どの程度まで記録を取る必要があるのか合意しておく。

1.2.17. 重要な過程(例:素点-尺度点の換算表の作成など)について、どれくらいの労力をかけて点検を繰り返すのか合意しておく。

1.3. リソース

1.3.1. 採点、分析、報告を効率的かつ適切に行うのに十分なリソース(費用、時間、人員)が利用可能であることを確かめる。

1.3.2. 各リソースについて、予備の体制が利用できるかどうか確認する(例えば、等化の専門家が等化を実行できないときに、誰が代わりに行うかを決定したり、答案用紙のスキャナーが壊れたときに他のスキャナーを設置したりするなど)。

1.3.3. 予備の体制を利用した場合には時間的な問題が生じることに注意する。重要な人員が突然不在になった場合に備えた対応策を考えておく。

1.3.4. チームの適切なメンバーに仕事を割り振る(採点や分析、得点報告を引き受けるのは誰か、過程全体に対して責任があるのは誰か、など)。テスティングを担当している専門家は、例えば、各段階に関わる人が必要なスキルを持っているかどうか見定めなければならない。また、必要条件や仕様を定め、どの程度まで自動化するかを決めなければならない。

1.3.5. 必要となる時間を定め、採点、分析、報告の各段階について実施計画を作成する。仕上げと得点報告の締切は実現可能なものとする。

1.3.6. 必要となるソフトウェアやコンピュータ、ネットワーク環境を特定しておく。著作権で保護されカスタマイズされたソフトウェアや、ノート PC、デスクトップ PC、メインフレーム、ディスクスペース、サーバースペース、帯域幅分析などである。

1.3.7. 作業スペースがどのくらい必要であるか（スタッフや受検者全員が作業するのに十分な広さがあり、部屋数、机、椅子などについても十分であるか）を判断する。

1.3.8. データを電子的に安全に保つために必要となる手段を定める。

1.3.9. 必要となる補助的な備品（手作業で採点するための正解番号、計算機など）を使用できるようにしておく。

1.4. 利害関係者の要求と期待

テスト得点を使用する者たち――受検者、親や家庭教師、教師やカウンセラー、テストを運営する人々（代理店が含まれる場合も）――は、採点や等化、報告までにかかる時間に関して明確な要求と期待を持っている。これらの要求と期待は合理的なものであるべきであり、また関係者間で理解し合うべきものである（テストの使用に関する ITC ガイドライン（2000）付録 B―テスティングの過程で関係者と締結する契約に関するガイドラインも参照）。

1.4.1 必要に応じて、関係者間――利害関係者、業者、受検者、顧客など――で、採点や等化、報告について責任のある専門家と協議の上で、協定を策定する。契約自体にときどき変更が生じることに留意する。

1.4.2. 問題が発生した場合、どのように対処し、どのように解決するかを決定する最終責任と権限が誰にあるか合意しておく。

例えば、多枝選択式の問題に正解がない、インタビュアーが非常に傲慢である、受検者が騒がしい環境によって妨害される場合など。また一つの選択枝だけが正しいように問題が作成されたのに、後に受検者が他の選択枝も正しいと示した場合など。

1.4.3. 得点の発表後にミスが発覚した場合にどうするか、あらかじめ定めておく。

1.4.4. 受検者に対して、提示された正解が正しいかどうか質問をしたり、自分の得点に異議を申し立てたりする機会を与える。あるいは、受検者に対して問題提起を行う機会を与え、それが確実に対応されるようにする。

1.4.5. テストの各項目の採点について、正当性を主張するのに用いることができる文書を用意しておく。

1.5. 専門スタッフと職場の雰囲気

採点、分析、等化及び報告の責任者は、採点、分析、報告の過程について必要なスキルと知識を持つ専門家であるようにする。また、スタッフは仕事についての必要なコンピテンシーを持っているようにする。複数の人々が関わる場合には、彼らが協調して働くことが大切である。それゆえ、新しく人を雇うときには、そのチームで調和して働くことができるかが重要な考慮事項となる。

1.5.1. 仕事の速さについて、個人に理不尽なプレッシャーを与えない。

1.5.2. 労働時間が過度に長くならないようにする。

1.5.3. リラックスしながらも、細部に注意しながら仕事を行うようにする（特に誤りの防止に関して)。平穏だが志気が高い職場の雰囲気は、高い規準を維持するのに最も有効である。

1.5.4. スタッフに対して専門的な能力開発を行ったり、場合によっては人間的な成長や社会的スキルの訓練も行ったりするなどの支援を行う（例えば、今年のデータを分析する準備として過去のデータを用いたシステムの検証に参加する機会を与えるなど)。

1.6. 品質管理の手続きの独立した監視

品質管理の手続きが順守されていることを監視し、全ての問題や過失が記録されることを確実にするための仕事に対して、専門家を一人以上（プロジェクトの大きさや複雑さによる）配置する。品質管理の監視は、採点や分析、報告の過程に関わる者とは別個に行うべきである。この監視は、全ての利害関係者と共同して、例えば、評価者内信頼性や、データ入力でのエラー率など、各過程を監査する目的で実行されるべきである。専門家団体が本ガイドラインを採用し、監視の過程において積極的な役割を担ってもよい。

1.7. ミスの記録と報告

1.7.1. テスティングに携わる者は全員、活動の様子及び発生した過失や問題の記録に関して合意された手続きに従う。

1.7.2. 各段階について誰が責任者であるか、あらかじめ合意しておく。

1.7.3. 全ての活動を記録する。定められたチェックシートを用いて各過程が実行され、チェックマークが付いていることが分かるようにする。

1.7.4. 全てのミスや過失について(原因が既に分かっていようとも)詳細に記録する。記録する内容は、ミスの内容、誰がいつ発見したか、その意味するところは何か、どのような対処がなされたかである。また、被害が生じる前に発見されたミスについても記録する。

1.7.5. 他の専門家にも、ミスについて適宜忠告する。そうした忠告のために時にはミス防止のための特別会議を開催する。

1. 7. 6.　今後のミスや過失をどのように防ぐかについて、文書にまとめる。

第 2 部：段階別のガイドライン

本ガイドラインは、採点、分析及び報告を実施していく上で取り上げられるべき手順を示す。大規模テスティングの実施においては、各段階はよく考えられ、注意して実行されるべきである。採点手続きを実際のデータで行う前に試行し、効率よく処理されるようにする。受検者が何千人もいる場合には、本ガイドラインに確実に従うべきである。受検者が数十人である場合には、一部簡略化して本ガイドラインの原則を実行するべきである。なぜなら、手続きのいくつかは多大なリソースを必要とし、大規模な集団を前提としているからである。そうした手続きは適宜、小規模な集団に柔軟に適合させるべきである。

2.1. 報告する内容

各手順を実行する前に、最終成果物である報告に関して合意しておくべきである。何を、どの程度詳しく、誰に、いつ、報告するかを決める必要がある。機関や受検者に得点を数値やその派生物（スタナイン尺度など）の形式で報告するだけでは不十分である。得点を適切に解釈することが非常に重要である。実際、テストの開発から採点、分析までの全ての段階において、最終成果物、すなわち報告される得点の解釈を考慮すべきである。この意味で、テスト開発全体の潜在的な目的、あるいは暗黙の最初のステップは、報告された得点に対する適切な理解が確保されるようにすることである。従って、得点解釈のさまざまな側面は最初から明確にしておくべきである。一つの得点だけでなく下位得点を報告することに関して、下位得点を報告するべきか、下位得点が利用されるのかなど、関連すること全てについて合意をしておくべきである。

2.2. 受検者について参考資料となるデータ

受検者についての参考資料となるデータや経歴に関するデータは、品質管理の過程において、受検者が本人であることを確かめたり、予想外の結果が生じたときに解釈したり、等化の目的でマッチさせるグループを作ったりする目的のために非常に有用である。次の手順が推奨される。

2.2.1. 法的に許されるのであれば、参考資料となるデータや経歴に関するデータ（年齢、ジェンダー、人種・民族、学歴、過去のテスト得点など）を集めておく。その方法には、事前申し込み時や、インターネット上での申し込み時や、テスト実施後に、受検者あるいは機関に申し入れる方法がある。収集するデータは、関連するデータに限られ、プライバシーについてはできる限り尊重する。

2.2.2. 可能であれば、受検者の経歴に関するデータを定期的かつ体系的に確認する。再受検者について矛盾が見られないかどうか注意を払う。

2.2.3. 参考資料となるデータと得点との間の期待される相関に関する研究を行い、現

在のデータの得点パターンと、他の情報、例えば過去のデータセットや研究成果等との不一致を探す。例えば、あるテストでは、大人の方が若者よりも成績が良いとする。研究の結果、そのようなテストでは若者の方が成績が良いことが期待されると示されたら、ミスが生じていないかどうか、採点過程を検証すべきである。

2.3. 採点

2.3.1. 全受検者の解答の入手と保管

解答用紙がある場合には、受検者の全ての解答用紙を保管し、適切な場合には各受検者の受検番号とともに電子的にも保存する。そのような資料（印刷物や電子情報）は、専門家の実務やその地域での法的な必要性を踏まえて、最短及び最長保存期間を定める。これには入手方法や保管方法によらず、個人が確認できる解答用紙や、解答や得点についての電子的な記録や、得点情報の記録が該当する。

2.3.1.1. 紙筆テストの解答用紙は、国やその地域などの法律に従って一定期間保存する（そのような法律がある場合）。

2.3.1.2. 電子的な記録に関しては、無停電電源装置（UPS）と、コンピュータやその他の機器の予備電源を用いて、「中断」されたり、データが失われたりすることのないようにする。

2.3.1.3. スキャナーを用いる場合は、定期的に確認し調整する。

2.3.1.4. スキャンした結果は、日常的に人が確認する。

2.3.1.5. 受検番号を厳密に運用するために、受検者のデータベースを確認する。例えば、重複して同じ受検番号が割り当てられていないか確かめる。

2.3.1.6. 全てのデータを安全に保護する。可能であれば、得点と個人を特定するような情報（名前など）とを分離し、個人情報を保護する。例えば、経歴に関するデータと得点に関するデータを別々のファイルにして、ID で統合できるようにする。これらのことは全て、データの安全やデータの保存に関する法律に従う。

2.3.1.7. 採点アルゴリズムが正確かどうか、換算表や基準が適切に使用されているかどうか確認する。

2.3.2. 客観式テストにおける採点

データが処理されて安全にデータベースに保存されたら、受検者の解答から通常は素点が計算される。例えば古典的テスト理論（CTT: classical test theory）では、通常、素点は正答した項目数であるが、当て推量に対する修正がなされたり、項目によって配点が異なったりする場合がある。項目反応理論（IRT: item response theory）では、素点とは潜在能力――しばしば“θ”あるいは“特性値”と呼ばれる――である。採点は、正解番号

の誤りなど、さまざまな種類の過失による悪影響を受ける。過失によって非常に低い得点になる場合もある。そうした過失を見つけるために、以下の品質管理に関する手続きを利用する。

2.3.2.1. データ構造が、データの記録形式の要件と適合するかを確かめる（例：ファイル内の項目の順番）。

2.3.2.2. 無効なデータを削除したり、欠損情報を記録したり、重複するデータに対処したりする際には、承認されたルールに従う。

2.3.2.3. サンプルデータについて、得点の範囲や記述統計量の値をテスト出版者の基準と比較する（提供されている場合には）。サンプルデータの統計は（サンプリングによる誤差分散によって期待されるよりも）いくらか逸脱してもおかしくはないが、差が大きい場合には注意し、場合によっては調査する。

2.3.2.4. 個人あるいは集団における極端な得点－低い場合も高い場合も－（紙筆テストであってもコンピュータによるテストであっても）を見直す。極端な得点は、得点の計算時のミス、受検者のいい加減な態度、データ取得時のミスのいずれかの問題の可能性を示唆している。

2.3.2.5. ある受検者について、互いに関連する下位テスト得点間の違いが予想されるよりも大きい場合には、その受検者のデータを見直す。そのための基準値をあらかじめ定めておく。

2.3.2.6. 項目分析を行い、項目統計量を調べる。――項目統計量を見ない限り、正解番号が 1 項目だけ間違っている場合を検出するのは難しい。(正解番号を間違えた項目は、難度の高い問題に見えることが多く、負の識別力を持つように見えることもある。例えば、基準と負の相関があるように見える。)

2.3.2.7. 各項目の無解答率を確認する。一部の受検者の採点から、ある項目が誤って除外されることがある。

2.3.2.8. 異なる条件で受検したグループには特別に注意を払い、データについて追加の確認を行う。例えば、異なる日付にテストを受けた集団や、異なる版のテストを受けた集団や、異なる解答方法を用いた集団などに対してである。

2.3.2.9. 例えば受検会場やテスト実施者、あるいは同じインターネット接続を用いたコンピュータなど、主な受検者集団の単位ごとに基本統計量を計算し、見直す。特定の会場に誤った版のテストを割り当てたなどの過失がありうる。

2.3.2.10. リソースが許すならば、別のチームもランダムに抽出した解答用紙を元に分析し採点するとよい。後に、チーム間で結果を比較することができる。

2.3.3. パフォーマンステストや面接等での評価

多枝選択式（MC; multiple-choice）項目の採点は客観的で（決められた正解番号に基づい

ており）信頼性が高い一方で、自由回答式(OE; open-ended)項目（パフォーマンスアセスメントや、自由回答形式のアンケート、作業事例テスト、ロールプレイなど）の採点は主観的なところがある。OE 項目はしばしば人間が採点し、その採点の影響を受けるため、MC 項目の採点よりも信頼性が低い傾向がある。しかし、OE 項目の採点にともなう主観性を減らしたり、採点の信頼性や正確さを改善したりするために、さまざまな方法がある。

2.3.3.1. パフォーマンスや作業事例テスト、ロールプレイ、インタビューの評価は、資格のある者、講習会や公式の教育を受けた者、あるいは必要な知識や経験を有する熟練した評価者が行うようにする。

2.3.3.2. OE 項目の採点指示書は明確で分かりやすく構成されたものとする。採点指示書を作成する助けとするために OE 項目のプリテストを実施すべきである。

2.3.3.3. OE 項目の各レベルの範囲を特定する作業(range-finding activities)を行い、ルーブリックの点数ごとに解答例を挙げる。採点の講習会では解答例を示す。

2.3.3.4. 評価者に対して、評価を行う前に講習会に参加するように求める。講習を受ける事によって採点指示書の内容をよく理解し、実際の受検者の解答の評価を任せられる前に評価の練習をすることができる。

2.3.3.5. 実際の評価作業に入る前に、講習会での練習に基づいて評価者の能力を評価する。

2.3.3.6. コストや使用できるリソースにもよるが、各アセスメントについて少なくとも二人が評価を行うようにする。

2.3.3.7. テストの重要性や分量、その他の要因にもよるが、(財政上の理由や他の考慮すべき事項により）受検者全員の評価を行う者が一人しかいない場合には、採点の信頼性を見積もるためにサンプル（例えばデータ全体の 10%）を二人で評価する。

2.3.3.8. OE 項目に対してコンピュータによる採点が用いられる場合には、人間の評価者がその採点を確認するようにする。実際の運用に入る前に研究結果に基づいてコンピュータによる採点の利用を正当化する。

2.3.3.9. 評価者は互いに独自に評価するようにする。

2.3.3.10. 採点過程の信頼性を評価するために、統計的な手続きを用いる。すなわち、偶然による相関係数の影響を除去した上で、評価者内や評価者間の一致度や不一致度を計算する。

2.3.3.11. 評価の質をリアルタイムで定期的にモニターし、フィードバックできるようにする。

2.3.3.12. 評価者が期待できない場合には（評価が信頼できなかったり、他の評価者の得点とかけ離れていたりする場合には)、本人に知らせて再トレーニングを検討する。問題が解決されない場合は、評価者を変えることを躊躇しない。

2.3.3.13. 評価者間の違いが大きい場合の方針を決めておく。違いが小さい場合には、平均を取るか、端数処理の問題を避けるために足し合わせる。違いが大きい場合には、経験を積んだ評価者が調整することになるだろう。

2.4. テスト分析

2.4.1. 大規模な多枝選択式及び自由回答式のテスティングのための項目分析

項目分析を行って基本的な統計量を計算し、項目の特徴や、その項目の点数を合計点に加えた場合にどのように機能するか確認する。受検者が少ない場合を除いて、全ての実施ごとに、また全てのテストの版ごとに項目分析を行うことを勧める。項目統計量には困難度（パーソナリティ検査の項目であれば、項目の“黙従傾向”）と識別度がある。テストの開発時に用いられたモデルによるが、多くの場合、各項目の IRT に基づくパラメタを計算することができる。さらに、項目分析により一般的なテスト統計量が得られる（信頼性や、標準誤差、平均、標準偏差、テスト情報量、受検者の反応の分布など）。次に示す手続きは、受検者数が少なくない場合には常に、用いられたモデルに応じてなされるべきである。

2.4.1.1. 項目分析には信頼できる方法を用いる。そのプログラムには適切な技術文書が備わっているようにするべきである。

2.4.1.2. 項目分析に用いるプログラムが良くないと思われる理由がある場合、あるいは新しいプログラムを用いている場合には、二つのプログラムを併用して、結果を比較する。

2.4.1.3. テスト実施後に項目分析を行う。あるいはテストが定期的に実施される場合には蓄積されたデータを分析する（例：実施後 3～5 年以内に）。（全てのデータが手に入る前に）部分的なデータに対して項目分析を行い、エラーがすぐに判明するようにする。

2.4.1.4. 受検者についての結果を出す前に項目分析を見直す。

2.4.1.5. 項目分析によって正解番号の誤りが判明することがある。例えば、選択する人がとても多い“迷わし”（誤答選択枝）が実は正解であったり、項目間相関が負の値になったのは、逆転項目の結果を反転させていなかったためだったりなど。ある項目に関する分析結果に問題がある場合には、正解番号や項目の中身を見直すべきである。

2.4.1.6. 正解番号を修正したり項目を削除したりした場合には項目分析をやり直す。そして採点表や等化の仕様などの文書を改訂する。

2.4.2. 新しいテストフォームや項目の等化とキャリブレーション

同じ版のテストを同時に受検した受検者内でのみ競争が行われる場合のように、等化が重要でない場合もある。等化をしないと、先行するテストの受検者の得点を、その後に実

施される新しい版の他の受検者の得点と比較することはできない。異なる版の得点を同一尺度上に載せる必要がある場合には、新しい版を等化して、古い版と比較可能となるようにする必要がある。等化の結果、全ての版のテストの得点が同じ意味を持つようになる。等化はテストの実施前にも（事前等化）、そして／あるいはテスト実施後にも（事後等化）行うことができる。等化は、項目レベル、尺度レベル、テストレベルのデータを用いて行うことができる。等化には異なる見方や方法がある（例：線形等化法、等パーセンタイル等化法、IRT に基づく共通受検者等化法や共通項目等化法）。

等化は、等化法やデザインに応じて、通常たくさんの受検者を必要とする。（Kolen & Brennan, 2004; Lamprianou, 2007 参照）

2.4.2.1. 等化に関して説明の付かない問題（例：得点が期待されたよりも低い）が生じた場合には、全ての版が同じ標準化された状況で実施されたかどうか確認する。実施状況が標準化されていなかった場合には、異なる状況であった事による影響を判断する。

2.4.2.2. 等化の手続きやデザインが正しく実行されたことを確認するための手続きを開発する。

2.4.2.3. 等化の手続きが前提としていることを調べる。あるいは異なる前提を持つ異なる等化の手続きでも同じような結果となるかどうかを調べる。共通項目のパラメタの安定性を確認する。等化のために共通項目のセットを用いているのであれば、そのセットからいくつかの項目を削除する場合の根拠、及びその決定が得点や等化された合格点に与える影響について文書にまとめる。スクリーニング後の共通項目のセットの内容と統計的な特徴についても文書にまとめる。この規準は共通受検者デザインにも適用されるが、その場合はスクリーニングの対象は受検者である。

2.4.2.4. 受検者の得点と、受検者の参考資料となるデータから予想される得点とを比較する（2.2.1.参照）。もしも乖離がある場合には、得点を再確認する。

2.4.2.5. 得点や合格率の比較を過去にわたって実施する。適切に開発された大規模なアセスメントでは、年ごとの変動は小さいことが多い。大きな変動は、テスト得点の等化における問題、例えば受検者集団の特徴の変化などを示唆する。

2.4.2.6. テストが複数回実施されるとき（受検者が多数で少ない回数の実施ではなく、受検者が少数で何回も実施されるとき）には、テスト得点の安定性を監視するための品質管理の方法を実施する。例えば、シューハート管理図や CUSUM 管理図、時系列モデル、チェンジポイントモデル、データマイニングの手法など（Von Davier, 2011 を見よ）。

2.4.2.7. 受検者を合格、不合格に分けたり、成績のレベル別に分けたりする場合には、その合否の割合や各レベルの割合を確認する。過去に行われた結果や受検者の

背景、似たようなテストから予想される割合と比較する。

2.4.2.8. 合格点の設定においては、委員会間で一貫性があるようにし、説明責任を果たせる方法を用い、その過程を記録する。標準的な過程から逸脱した場合についても記録する。

2.4.2.9. 紙筆テストとして開発されたテストをコンピュータで実施するなど、実施形式を変える場合には、新しいテストの特徴をそれまでの結果と比べる必要があり、場合によっては新しいテストを古いテストへ等化する必要がある。

2.4.2.10. ハイステークスなテストでは、等化の結果を独自に再現できるようにできる限り努力し、その過程には第三者も含むようにする。

2.4.3. 標準化得点の計算

多くの場合、標準化得点は得点を解釈するのに役立つ。どのような標準化得点が用いられるにせよ（例：スタナインや十分位など）、標準化得点の計算には素点が用いられる。尺度得点や標準化得点、パーセンタイル得点の計算には、パラメタや換算表が用いられる。通常、素点（正答数あるいは当て推量を調整した正答数）や θ 値（IRT に基づくテストの場合）はそのテストの尺度上に変換される。換算は、参照テーブルや関数（例：線形変換）によってなされる。

2.4.3.1. 尺度得点を得るために、素点を適切に換算する。

2.4.3.2. 正確を期すために、尺度得点の換算とその手続きを確認する。

2.4.3.3. 正しい換算方法が用いられたことを確認する。

2.4.3.4. 素点が低ければ標準化得点も低く、素点が高ければ標準化得点も高くなっていることを確かめる。

2.4.3.5. 換算後、追加の手続きを行うべき場合もある（例：各回で報告する最低点と最高点をそろえる）

2.4.3.6. 換算表やパラメタについて、テストの新しい版と他の版とを比較し、目立った相違や類似があるかどうか見る。

2.4.3.7. 尺度上に現れた経時的な変化について説明する。

2.4.3.8. いくつかの得点は人の手で計算をして、コンピュータが計算した結果と比較する。

2.4.3.9. 散布図を描いて、素点と標準化得点の間の統計的な関係を確認する。

2.4.3.10. 二つの異なるコンピュータプログラムを用いて標準化得点を計算し、それらを比較する。

2.4.3.11. 技術マニュアルなどに、素点を標準化得点に換算する手続きの詳細を記述しておく。この方法はテストの版ごとに異なるので、手続きも版ごとに記述されるべきである。

2.4.4. テストの安全性の確認

得点が報告された後に不正行為が発覚すると、テストやテスティングシステムのセキュリティや品位を傷付ける深刻な問題となる。残念なことに、不正行為は監視や他の抑止策を用いても完全に防ぐことはできない。不正行為への誘惑は、特にハイステークスなテストにおいては、非常に大きくなりうる。不正行為へのたゆまぬ対策として、弁護士らにセキュリティチェックを見直して、その妥当性を確認するよう相談すべきであろう。国レベルでのハイステークスな教育テストでは、不正は個人レベルだけでなく、学級や学校、地域、職場単位で起こる。テスト会場や携帯電話あるいはインターネット上のウェブサイトを通じても起こる。雇用の分野では、自宅から（インターネットを介して）仕事を得るためのテストを受けることが多くなっているが、なりすましのリスクやさまざまな形態の偽装が増加している。セキュリティチェックの副効用としては――不正を割り出すのとは別に――テスト実施やデータの収集・保管における問題の指標にもなりうることである。次のような予防措置を講じるとよいだろう。

2.4.4.1. 受検者の身元は、受検会場にいるときか自宅でテストをしているときに、写真付きの身分証か、虹彩スキャンや指紋などの生体認証機能を用いて照合する。遠隔地にいる受検者を同定するために進んだ技術を用いる。

2.4.4.2. 複数のテストフォームを用いる方がよい。テストフォームが一つの場合は、知り合い同士の受検者（例：近所に住んでいたり、同じ学校に通っていたり）を近くに座らせない。例えば、名前順に座らせるようにする。

2.4.4.3. 座席番号を記録し、誰がどこに座っていたかのリストを作り、他人の答案の書き写しが行われたかどうかを分析する際の助けとする。

2.4.4.4. 必要に応じて（例：書き写しが疑われるとき）、書き写しが行われたかを見るための統計的な指標、すなわち、受検者の解答用紙と同じテスト会場の他の受検者の解答用紙の類似度に基づく指標を用いる。

2.4.4.5. 訓練された信頼できる試験監督を雇い、彼らを定期的に監視する。試験監督が利益相反を持たないことを確認する。

2.4.4.6. 異常な、あるいは予想外の反応パターンをチェックする（例：難しい項目に正答し、易しい項目に間違えるなど）。

2.4.4.7. 試験前と試験中に受検者の筆跡のサンプルを入手し、なりすましや疑わしい場合に確認する際の助けとする。この手続きは、身分の確認が問題でない場合には省略してよい。

2.4.4.8. （繰り返し受検が可能なのであれば）繰り返し受検した場合の得点の不一致を、現在と過去の得点との合理的な差を示す統計的分布を用いて分析する。差が大きい場合には、その受検者のなりすましを示唆しているか、あるいはテスト実

施前に項目に関する情報を得ていたことを示唆する。もちろん、その他の説明として、同じ、あるいは類似のテストを受検する事による練習効果も考えられる。

2.4.4.9. 不正が疑われる人の取扱いについて(必要であれば法定の)文書で定めておく。不正や不正行為に対する方針が施行されており、監視が行われていることを全受検者に事前に知らせておく。

2.4.4.10. 標準化されたテストでは、生徒の得点を上げることに関心のある教師がいる。そのため、教師が標準化されたテストの得点にアクセスできるようにするべきではない。

2.4.4.11. テスト用具とその結果を安全に守るために鍵の掛かったキャビネットや安全なサーバーを用いる。テストやその項目の開発に関わる全ての人が信頼でき、セキュリティに関して十分概要を知っているようにする。テスト項目の機密性についても、最初から最後まで、素案の段階から確保されなければならない。項目は、安全な方法で業者とテスト開発者の間でやり取りされなければならず、全てのファイルは、権限のない人々が容易にアクセスできるような PC やサーバーではなく、USB メモリやスタンドアローンのノート PC 上に保管され、作業が行われる必要がある。

2.4.4.12. テスト用具の保存や送信ができないように、テストの提示に用いられるコンピュータをロックダウンする必要がある。テスト用具を送信できる場合には、インターネットへの接続は避けるべきである。

2.4.4.13. テスト用具を秘密にしておくために、(カメラや携帯電話で) 写真を撮られないようにする。

2.4.4.14. 全受検者を公平に取り扱うために、受検者の機密性は、テストの受検及び採点の全段階において守られなければならない。

2.5. 報告

2.5.1. 得点の報告

得点は受検者とテスト使用者（顧客）の両方に公表される。理想的には、得点報告書は印刷可能な形で提供されるべきである。テストによっては、インターネットが標準的な得点を報告する方法になっている場合もある。報告は、得点の意味が受検者にも顧客にも明確であるような方法で行われなければならない。

2.5.1.1. 得点報告書やその解釈のための手引きが、分かりやすく教育的なものとなるように、受検者のフォーカスグループを利用するか、場合によっては「発話思考法」「実験研究」「1 対 1 のインタビュー」を行い、情報を集める。

2.5.1.2. 得点を受け取った誰もが、その得点を解釈するための指導を受け、テスト得点

を適切に理解できるようにする。テスト使用者が正当な解釈を行っているという証拠を用意しておく。

2.5.1.3. テストの説明や得点の意味を記載した報告書をコンピュータで作成し、専門的な内容を分かりやすく説明し、報告を受け取った人に分かりやすいようにする。

2.5.1.4. 国際的なテストや全国あるいは地方レベルのテストにおいて必要であれば、テスト結果をその都度取り出せるような中間的データ保管システムを用いる。

2.5.1.5. 得点によってどの程度信頼していいのかを明確にする（例：ハイステークスな決定を行うには下位得点の信頼性が低すぎるなど）。下位得点を報告するかどうかの決定は、(a)テスト理論や(b)テスティングの目的及び下位得点の心理統計的な特徴にも基づくべきである。

2.5.1.6. 結果や得点がメディアや政治家に報告されるときには広報活動の専門家の助けを借りる。

2.5.2. 得点報告の安全性を維持するための対策

2.5.2.1. 個人の得点報告書が受検者によって偽造されないように対策を講じる。

2.5.2.2. 機関への報告書を改編することは避ける。改編は深刻な問題を生じうる。得点を変更する必要がある場合には、指定されたソフトウェアを用いるか、報告書を作り直す。

2.5.2.3. 保管や運搬のために、得点報告書の電子ファイルを暗号化する。

2.5.2.4. 得点報告書は適切な人にだけ送り、必要以上にいろいろな人に送らないようにする。全受検者に対して同じ報告書を送る方が容易であるが、受検者の機密性を守るために、関連する結果だけを各受検者に送るべきである。

2.5.2.5. 機関には、機関に直接送られた報告書だけが――（偽造されうる）受検者への報告書のコピーではなく――公的な目的で使われるべきものであることを知らせておく。また、機関には、送付された報告書を必ず検証するよう推奨する。

2.5.3. 記録

採点過程の全体（平均、標準偏差、中央値、値域、信頼性などの重要な記述統計量も含めて）を必ず記録する。また、現在の受検者群をこれまでの受検者群と比較する。これらのことをテスト得点の公開前か公開直後に終わらせておく。記録する習慣を付けておくことは、将来の各過程がより信頼でき正確なものになることにつながる。得点に関する情報を公開することは、採点、分析、報告の過程全体をコントロールする方法の一つとなる。次のことが重要である。

2.5.3.1. 重要な統計量や受検者群の比較など採点の過程全体について所定の記録を取ると同時に、テスティングの過程全体について段階ごとに（内部報告として）記

録を取る。

2.5.3.2. 古い版の記録が終わってから、新しい版が実施されるようにする。

2.5.3.3. 記述統計量、例えばジェンダーによる違いや年ごとの分布などを集積し、それらの統計量に誰もがアクセスできるようにする。それらの統計量に関する簡単な説明も載せるべきである。集積された統計量は、各受検者の匿名性を守る。

参考文献

AERA/APA/NCME. (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

Allalouf, A. (2007). Quality Control Procedures in the Scoring, Equating, and Reporting of Test Scores. Educational Measurement: Issues and Practice, 26: 36-43.

Bartram, D., Hambleton, R.K. (Eds.) (2006) .Computer-Based Testing and the Internet. West Sussex: John Wiley & Sons.

Cizek, G. J. (1999). Cheating on tests: How to do it, detect it, and prevent it. Mahwah, NJ: Lawrence Erlbaum.

ITC (2001). International Guidelines on Test Use. International Journal of Testing, 1: 95-114.

ITC (2006). International Guidelines on computer-based and Internet-delivered testing. International Journal of Testing, 6: 143-172.

Kolen, M. J., and Brennan, R. L. (2004). Test equating, linking and scaling: Methods and practices. New York: Springer.

Lamprianou, I. (2007). Comparability methods and public distrust: an international perspective. In Newton, P., Baird J., Goldstein, H., Patric, H., & Tymms, P. (Eds.) Techniques for monitoring the comparability of examination standards. Qualifications and Curriculum Authority, London.

Nichols, S. L. & Berliner, D. C. (2005). The inevitable corruption of indicators and educators through high-stakes testing, Educational Policy Studies Laboratory, College of Education, Arizona State University.

Rhoades, K., & Madaus, G. (2003). Errors in standardized tests: A systemic problem. (NBETPP Monograph). Boston, MA: Boston College, Lynch School of Education.

Texas Education Agency, Pearson Educational Measurement, Harcourt Educational Measurement & Beta, Inc. (2004) Chapter 9: Quality control procedures. Texas Student Assessment Program. Technical Digest (2003-2004) http://www.tea.state.tx.us/student.assessment/resources/techdig04/

Toch, T. (2006). Margins of error: The testing industry in the No Child Left Behind era. Washington: Education Sector Report.

Von Davier, A. (2011) Statistical Models for Test Equating, Scaling, and Linking. Springer

Wild, C. L., & Rawasmany, R. (Eds.) (2007). Improving testing: Applying process tools and techniques to assure quality. Mahwah, NJ: Erlbaum.

Zapf, D. & Reason, J. (1994). Introduction: Human Errors and Error Handling. Applied Psychology: An International Review, 43: 427-432.

（日本語版　2017 年 4 月 5 日）