第1章
効果的なテスト作成に至る12のステップ

Steven M. Downing
（イリノイ大学シカゴ校）

効果のあるテストを作るには，正しい理論に基づいた教育測定の原理を基礎に系統だった細かいアプローチが必要となる。この章では，多くの学力，能力，技能テストの作成でなさねばならない12の代表的なテスト作成手順を個々に論じている。これら12のステップから成る効果的なテスト作成法に従えば，解答選択式，解答構築式どちらのテストの場合でも，狙いとするテスト得点の解釈が妥当であるという証拠を最大にするような結果が示せるであろう。

これらの12ステップは1つの枠組みとして提供されている。それはテスト作成者にとって，通常作成に伴う多くの課題をまとめるのに役立つに違いない。それはステップ1での細かい計画から始まり，内容の定義をめぐる議論から輪郭，テスト刺激（問題項目，作業課題など）の創作，実施，採点，報告，そして記録と，すべての重要なテストの開発活動を含んでいる。それに関連する『スタンダード』も各ステップで引用され，全体を通して妥当性の議論に光を当てている。この章は『テスト作成ハンドブック』の内容の概括で，12ステップの1つ1つを他の章と関連させながら述べている。

効果のあるテストを作るには，テスト得点からなされる推論を支持するのに十分な妥当性が確保できるように体系的で，うまく組み立てられた方法が必要である。通常，そこには「テスト開発」とか「テスト構築」とかという言葉で呼ばれる大小無数の細かい議論から成る一大事業がある。これらすべての詳しい1つ1つはテストが測ろうとしている内容領域で，受験者の学力や能力を公平に，また一貫して推定できるテストが生まれるようにうまく実行されなければならない。それはまた，テスト得点が推測しようとしたことを支持する証拠となる記録が提供できるものでなければならない。

この章では，体系的なテスト作成のモデルを12の個別課題または活動にまとめて論じている。この章で概略述べられているすべての活動や課題の詳細はこの本の他の章で詳しく述べられる。したがって，この章で述べることは『テスト作成ハンドブック』の内容を概観し，その入門になると考えられるものである。この章ではテスト作成に結びついた重要な議論の1つ1つを網羅的に論ずることは考えていないし，それらの活動から得られる妥当な証拠のすべてを述べることも考えていない。むしろこの章では，テスト作成者にとって潜在的に役立つと考えられるモデルまたは型板を提供する。それに従い，1つ1つの重要なテスト作成の活動によって，関心の対象である構成概念の効果的な物差しを創りだす可能性が最も高くなるように十分な注意が払われることになるであろう。

この課題と活動を12の個別ステップにまとめ上げることはある程度恣意的である。ここで挙げる課題はもっと少ないか，あるいはもっと多くの個別ステップになるように違ったまとめ方ができるかもしれない。各ステップはある程度の細かさで，どのタイプのテストでも達成できるものでなくてはならない。問題形式が解答選択式（例：多枝選択式）であっても，解答構築式（例：短答式や論述式）であっても，あるいは作業式（例：高精度のシミュレーション）であっても成り立つもの，またテストの実施形態が伝統的なペーパーテストであろうとコンピュータ上で行うものであろうと使えるものでなくてはならない。各活動の強度や技術の高さは，作成されるテストの型や目的，そこで意図

される推論，あるいはテスト得点がもたらす影響度，テスト作成者の人材と技術の訓練度などによる。しかし，この章で記されるすべての課題は，どのテスト作成プロジェクトでもある程度の細かいレベルで実行されなければならないものばかりである。

これらの12ステップは，1つのテストプログラムにおいて，あらゆる妥当性の証拠集めや報告の材料として便利で組織立った枠組みを提供するものである。同時にそれは，テスト作成にかかわるものとして『教育・心理検査法のスタンダード』（米国教育研究学会［AERA］，米国心理学会［APA］，全米教育測定協議会［NCME］，1999）に関連して検討事項をまとめ上げるのに便利な方法となるであろう。これらの各ステップは，テストのすべての重要な活動と結果を要約する技術報告書の中に記録されるような妥当性の証拠を組み立てる上で，1つの重要なまとめ役と考えることができる。これらの各ステップはまた，テストの正確な目的，テスト得点の結果，そしてテスト得点から推論される望ましい解釈など，上記『スタンダード』（AERA, APA, & NCME, 1999）の中の1つや2つに関連してかなり適用できるものである。

表1.1のリストはテスト作成の12のステップを挙げたもので，そこでの課題，活動，そして問題点の簡単な要約である。そして，それに関連する『スタンダード』（AERA, APA, & NCME, 1999）の各条項が表1.1の各ステップごとに記されている。これらのステップは段階の最初から最後に至るまで直線的に，あるいは継続的に並べて個別にリストされている。しかし実際には，これらの活動の多くは同時に起ることもあれば，それらのステップの順序が多少修正されることもあるだろう。例えば，テストに求められる合否分割点とか体系的な規準設定活動などは，表1.1に示されているステップ9よりももっと早い段階で起こってくるかもしれないし，ステップ11での項目バンキングの議論も実際のテストプログラムでは，テスト作成順のもっと早い段階で起こるであろう。しかし，これらの活動の多くは他の活動の先行要件である。例えばテスト内容の定義付けは問題作成やテストの組み立てより前に起こるはずであるし，それがためにステップを順序付けておくことは多少恣意的であるとはいえ，意味のあることである。

この章に含まれる情報の大半は，どれがうまくいき，どれがそうでないか，長年の経験と学習で，また毎日の実際のテスト作成の中で基礎付けられるものである。ここで扱う研究文献はテスト作成の最もよい方法として支持されてきたもので，心理測定学や教育測定学の広範囲な領域から取られたものである。またMillman and Greene（1989）の章や『教育測定学第4版』（2006）のSchmeiser and Welchの章を読むと，この12のステップがテスト作成にとって効果のある有効なものであることを鼓舞してくれる。

ステップ1　全体計画

どのテストプログラムも始めるには何らかの全体プランが必要である。初めに決めなくてはならないのは，測ろうとするものの構成概念は何か，テスト得点についてはどのような解釈を望んでいるのか。計画しているアセスメントにとって最も適切なテスト形式は何か。解答選択式（selected response）なのか解答構築式（constructed response）なのか，あるいは作業式（performance）か，それともそれらの組み合わせか，決めなくてはならない。さらに，実施するテストの形態は，ペーパー式テストかそれともコンピュータ式テストか。それはどの程度まで正確な形で決めておかなくてはならないか。プログラム作成や作業の開始はいつから始めるのか。一連の課題はどんな順序で完成しなければならないのか。どの課題は他のどの課題を行ってからするのがよいか。それにはどれくらいの時間が必要なのか。それぞれ特定の課題の実行責任はだれが持つのか。その他いくらでもある議

表 1.1

効果のあるテスト作成に至る 12 のステップ

ステップ	テスト作成の課題例	テストスタンダードの関連項目例
1. 全体計画	全テスト作成活動についての体系的ガイダンス：構成概念；望ましいテストの解釈；テスト形式；妥当性を示す証拠の主な資料；明確な目的；望ましい推論；心理測定のモデル；作成日程；安全保持；品質管理	規準条項 1.1 規準条項 3.2 規準条項 3.9
2. 内容の定義	内容領域／母集団からの標本抽出計画；評価目的に関連した諸方法；内容関連妥当性の証拠を示す基礎資料；構成概念の描写	規準条項 1.6 規準条項 3.2 規準条項 3.11 規準条項 14.8
3. テストの仕様	内容の操作的定義；内容領域の抽出に関連して，体系的に説明できる妥当性の証拠を示す枠組み；集団規準か目標基準か？；望ましい項目特性は？	規準条項 1.6 規準条項 3.2 規準条項 3.3 規準条項 3.4 規準条項 3.11
4. 問題作成	効果的な質問の作成；テスト形式；証拠中心の原理に忠実な妥当性の証拠；問題執筆者と検閲者の訓練；効果的な問題編集；問題の不備からくる CIV	規準条項 3.6 規準条項 3.7 規準条項 3.17 規準条項 7.2 規準条項 13.18
5. テストデザインと組み立て	テスト形式のデザインと創作；特定形式のテスト問題の選択；計画された青写真による抽出作業；事前テストの配慮	規準条項 3.7 規準条項 3.8
6. テストの制作	出版活動；印刷または CBT 化；安全保持問題；品質管理についての妥当性問題	該当なし
7. テストの実施	標準化に関する妥当性問題；ADA 問題；試験監督；安全性問題；時間設定問題	規準条項 3.18 規準条項 3.19 規準条項 3.20 規準条項 3.21
8. テスト解答の採点	妥当性の議論：品質管理；正解の妥当性確認；項目分析	規準条項 3.6 規準条項 3.22
9. 合格点の設定	説明できる合格点の設定；相対基準か絶対基準か？；分割点の妥当性問題；規準の比較可能性；得点尺度の恒常性の維持（等化，リンク付け）	規準条項 4.10 規準条項 4.11 規準条項 4.19 規準条項 4.20 規準条項 4.21
10. テスト結果の報告	妥当性の議論：正確さ，品質管理；時期の適切さ；意義；誤用問題；チャレンジ出願；再受験	規準条項 8.13 規準条項 11.6 規準条項 11.12 規準条項 11.15 規準条項 13.19 規準条項 15.10 規準条項 15.11
11. 問題項目バンキング	安全性問題；有用性，柔軟性；効果的な項目バンキングの原理	規準条項 6.4

12. テストの技術報告書	体系的で綿密で詳細に記録された妥当性の証拠；12ステップの構成；推奨できること	規準条項3.1 規準条項6.5

略語：ADA，米国障害者法；CBT，コンピュータによるテスト；CIV，構成概念と無関係な分散

論，決定，作業課題，操作上の詳細など，テストプログラムのすべての面をどのように品質管理するのか。ステップ1，すなわち全体計画は，テスト作成プログラムに結び付いた全体の主たる活動のすべてについて体系的な枠組みを考えることである。それは頭で考えた最も重要な決定の大部分を表面に出し，プロジェクト全体に現実的な時間設定を与え，そこから生ずるテストの安全性と品質管理問題の大切さを最初の段階から強調することになる。

テストプログラムで最も基本的な決定は，正式のテスト作成活動が始まる前に作らなければならない。これらの基本的決定の1つ1つは，明解な理由と最終的なテストプログラムから作られるテスト得点についての主たる妥当性の証拠となるよりどころを示すことになる。

ステップ1の形に含まれる課題と決定の例は，計画されたテスト目的が明解で簡潔な形でよく描写される。テスト形式の目的は提案されたテストの操作的定義を形作り，テストの作成活動に関係するほかの妥当性関連のほとんどすべての決定を方向付けてしまう。究極的には内容の定義とテストの内容領域を定義するのに使われる方法，そして試験によって測られると想定されている構成概念などを決める主なステップが，すべてテストの明確な目的で述べられたものと直接結び付いている。心理測定モデルの選択で古典的な測定理論を選ぶのか，項目応答理論を用いるのかは，提案しているテストの目的に関係するかもしれないが，同時にそれは提案されたテストデータの使用法とテスト開発者と利用者の技術的高さにもよる。例えば，はっきりと述べられたテストの目的があって，よく定義された一連の教授過程やカリキュラムがあり，そこでの学生の達成度を査定することであれば，提案された構成概念やテストの内容を選ぶのに使われる方法，そして使用する心理測定モデルなどはテスト開発者により理にかなった選択ができるであろう。同様に，もしテストの目的が全国的で高度に競争的な専門教育プログラムがあり，そのための学生を選抜する能力推定であるならば，テスト得点からなされる推論が明解に表現され，関心のある主な構成概念が明解に描写され，内容を定義する方法，心理測定モデル，その他のテスト開発のために行われる主な決定がうまく導かれることになるだろう。

テスト作成全体を概括することの一部として，ほかにもなされなければならない基本的な決定には次のようなものがある。解答選択式テストの問題項目や作業式テストの作業課題や刺激をだれが作るのか。新しく書かれた問題項目や作業課題，その他のテスト刺激をだれが審査するのか。問題項目や作業課題の制作過程をどのように管理し，どのような時間設定で行うのか。また，テスト項目や作業課題の最終的な選択はだれが責任を持つのか。だれがテストを制作し，出版し，印刷するのか。一連のテスト制作過程を通して安全性の確保はどのように行うのか。すべてのテスト材料の正確さを保証するための品質管理はどのようになっているのか。いつ，どのようにしてテストを実施するのか，また，だれによってか。テストは伝統的なペーパー式テストか，それともコンピュータ式のテストか。また，求められれば，合否分割点や合格点をどのように決めるのか。またどのような方法で，だれがテストの採点を行い，受験者にそれをどう伝えるのか。だれがテスト項目や作業式テストの課題の安全性を保ちながら，項目バンクや項目プールを維持管理するのか。すべての主要な作業が完了するためにテスト開発に課せられた最終期限はいつか。データの結果，テストの評価というすべての重要な活動を記した完全な記録の作成はだれが責任を持つのか。

多くの重要な方法の中で，ステップ1はテスト開発の12の課題のうちで最も重要なステップである。初めがよいプロジェクトは終わりもよいことが多い。テストのこの厳しい開始段階はテストプロ