水本 篤 (2015).「LMS を利用した教育効果の測定」 大澤真也・中西大輔（編著）
『e ラーニングは教育を変えるか』(pp. 165–179). 海文堂出版

1. 目的に応じた研究手法の選択のために

2004 年に『英語教師のための教育データ分析入門 ― 授業が変わるテスト・評価・研究』（前田・山森 編著）という本が出版されました。この本では，英語教育学研究で使用されている主要な量的分析が初学者にもわかりやすく説明されていており，当時，大学院生だった私も常に参考としていました（この本がなければ私は博士論文を書けなかったと言っても過言ではありません)。そして，この名著を参考にしながら博士課程を修了した若手研究者を中心として，2012 年に『外国語教育研究ハンドブック ― 研究手法のより良い理解のために』(竹内・水本 編著）を出版しました。『外国語教育研究ハンドブック』では,『英語教師のための教育データ分析入門』のコンセプトを継承し，初学者（特に大学院生）でもわかるような書き方を心がけました。また，この 2 冊の書籍が出版されるまでの 8 年の間に新しく用いられるようになった効果量やメタ分析なども紹介し，実際にデータを触りながら分析を自分で分析に対する理解を深める目的で，コンパニオン・ウェブサイト（http://mizumot.com/handbook）では，MS Excel，IBM SPSS，R，そしてその他の統計解析ソフトウェアでどのように操作すればよいかという説明も行っています。さらに重要なこととして，外国語教育研究分野では，研究手法のパラダイム（分野における認識の枠組み）にも変化が起きていて，以前までは量的アプローチがほとんどであったものが，質的アプローチも広く用いられるようになってきています。そのような背景から『外国語教育研究ハンドブック』では，量的研究だけではなく，質的研究のセクションも設け，その基盤となる科学観やよく用いられる分析方法を詳しく説明しています。このように，研究手法も多様化している中で，LMS を利用した教育効果の測定を行うためにも，目的に応じた研究手法の選択が必要とされる時代になってきているといえるでしょう。

図 1 は『外国語教育研究ハンドブック』で示した，柳瀬（2006）に基づく，外国語教育研究の立ち位置です。図中の左側は「科学」となっていて，第二言語習得研究（SLA: Second Language Acquisition）などでは，量的なアプローチで客観的に研究を行うことが多く，外国

語教育の分野では，そのような研究がほとんどでした。しかし，外国語教育研究は，図中の右側にあるような実践も研究対象に含んだ分野であるため，実践と融和性が高い質的なアプローチの研究も近年，国内外で増えてきています。

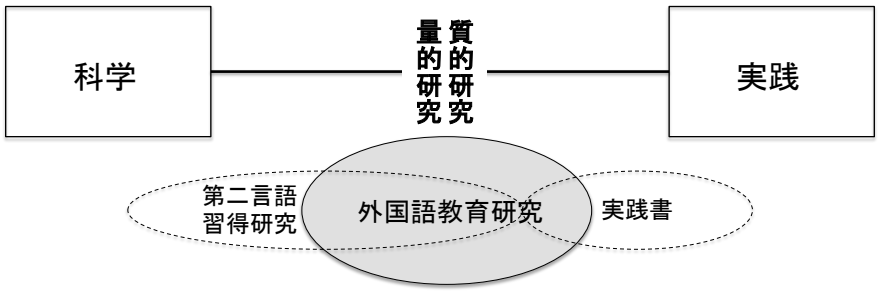

図 1.　外国語教育研究の立ち位置（竹内・水本, 2012, p. 8）

例えば，動機づけ（motivation）の研究では，これまで，質問紙を用いて，数値化したデータを基に量的アプローチで行うものがほとんどでした。しかし，質問紙調査でわかることは，その質問紙で想定されている概念（構成概念）のみになり，定義された動機づけの説明・理解だけになります。しかし，多くの教員が本当に知りたいことは，「どうすれば，学習者の動機づけが高まるか」というようなことであるため，そもそも知ろうとしている内容と，測定していることがずれているのです。磯田（2007, p. 4）は，この点について以下のように述べています。

> 動機づけ研究の目的が行動の「理解」にあるとすれば、意欲を高めるというのは行動を「変容」することにあると言える。行動を変容するためには、行動の背景を理解することはもちろん必要であろうが、理解するだけでは変容のための具体的な方策についての答えは出ない（p. 4）。

このような研究目的と手法のミスマッチが起こる理由の一つに，外国語教育研究の分野として量的アプローチばかりの実証主義的な論文が多いため，本当は質的アプローチを含めると研究手法の多様性があるにもかかわらず，量的アプローチ以外の認識論が理解されていない（もしくは理解しようとしない）ということが考えられます（高木，2014）。実際，Mizumoto, Urano, and Maeda (2014) は，全国英語教育学会の学会誌である *ARELE* (Annual Review of English Language Education in Japan) の第 1 号から第 24 号（1990–2013）に掲載された論文 413 本のうち，82.6%（341 本）が量的研究であり，質的研究はわずか 3%（13 本）しかなかったことを報告しています。論文を執筆し，ジャーナルに投稿する場合には、その分野での先行研究やその先行研究で用いられている分析方法を踏襲するというのが，もっともよく使われるストラテジーであるということを考えても，質的アプローチの論文が該当分野のトップ・ジャーナルで掲載されていない場合は，そのような方法を実際のデザインに落とし込むことができず，論文にもなりづらいというサイクルが考えられます。ただし，国内外で，これまでの量的・実証主義的なアプローチのみを良しとする考え方が見直されつつありますので，今後は，質的アプローチの論文も増えることが予想されます。

ここで大切なことは，「量的アプローチのほうが，手本となる先行研究が多いので論文が執筆しやすい」とか，「質的アプローチや，量的と質的を合わせた混合研究法（mixed methods）のほうが現在の流れでは掲載されやすい」とか，そういうことではなく，「質的アプローチの認識論にも造詣を深めながら，目的に応じた研究手法の選択していけるようにする」ということです。ただし，量的，質的アプローチは元々，依拠する科学観が違うため（住，2012），「その選択は研究者の関心によって相対的に決定されるものである」という科学観に基づいた構造構成主義（西條，2005）を『外国語教育研究ハンドブック』では推奨しています（図 2）。構造構成主義によって「共通了解可能な知識を獲得する」（住, 2012, p. 249）という意味での一般化が可能になり，「量的 or 質的」という対立した考え方ではなく，同じ次元で目的に応じて研究手法を選択することができます。そのような広い視野から研究対象に向き合うためにも，量的アプローチ以外の認識論も理解し，研究者自身の引き出し

を増やしていくように努めなければなりません。

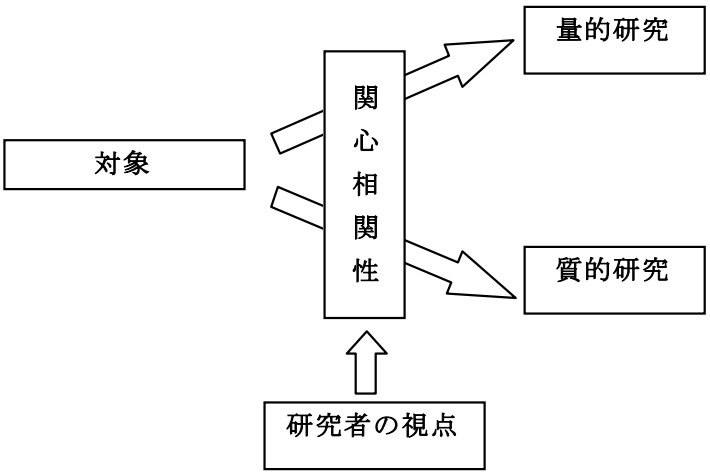

図 2.　構造構成主義における量的研究と質的研究の位置づけ（住, 2012, p. 248）

具体的には，教育効果を「測定」（measurement）という側面で考えると，従来の量的・実証主義的なアプローチでは，テストや質問紙（アンケート）を教育介入の前後で実施し，結果を数量化することで，その教育効果の検討を行うという方法が，現在でも，もっとも多く行われています。しかし，測定もその概念に含む「アセスメント（評価と訳されることもあります)」（assessment）を教育効果検証の対象とした場合，アセスメントはテストや質問紙の結果だけではなく，学習者がどのようなことができるかという証拠を系統的に収集する過程となります（Bachman, 2004)。そのため，ポートフォリオ，観察，ジャーナル，パフォーマンス，自己評価，学習者同士の相互評価など，学習プロセスに焦点を当てた新しい評価（alternative assessment）を用いることも可能です。このようなアプローチは，教育実践の観点からも好ましいことがわかります。指導後にのみテストや質問紙で測定を行うようなものは，Assessment of learning（学習の評価）ですが，学習プロセスに焦点を当てる新しい評価は，Assessment for learning（学習のための評価）と呼ばれ（Stiggins, 2002)，近年では後者の重要性が認識されています。前述の動機づけ研究の例のように，テストや質問紙では，測定していない概念以外の学習効果を調べることはできません。しかし，（LMS を

利用した）教育効果の測定においても，学習プロセスに焦点を当てた新しい評価，そして質的アプローチを用いることによって，量的アプローチでは見えなかった教育効果が見えてくるかもしれません。そのため，教育効果の検証を研究対象として考える場合には，さまざまなアプローチを考慮しつつ，目的に応じて研究手法を選択するべきです。

ただし，新しい評価や質的アプローチを用いた教育効果の検証においても，やみくもにデータ（証拠）を収集するのではなく，研究として成立させるためには，何を研究対象としたいかという点については，データを収集する前に考えておくべきです。そうすれば，自ずとどのような分析方法が必要とされるかわかるため，分析方法を考えずにデータを収集してから，自分でもどうすべきかわからないので誰かに相談するというようなことはなくなるはずです。

「教育効果の測定」というテーマでは，外国語教育研究の分野のみならず，教育心理学など様々な分野で研究が行われているため，研究対象や必要とされる手法は先行研究を概観することで確認することができます。現在では，Google Scholar（http://scholar.google.com）や研究機関で提供されているデータベースで多くの先行研究を確認することが可能です。また，専門書でも初学者向けに日本語で書かれた文献が多く存在しますので，そのような書籍で主要な先行研究を知るきっかけを得てもよいでしょう。

前述したように，現在，量的・実証主義的なアプローチの論文が先行研究でも多いはずですが，量的なもので興味のある内容が先行研究にあれば追試（replication）を検討すべきです。これまでは，「独創的な研究」が求められ，1 度だけの実験で終わるというものがほとんどでした。しかし，分野として，量的アプローチでは再現性を重視するために，追試を推奨するという風潮になってきています（国際誌の *Language Teaching* では 2008 年から Replication Studies のセクションが追加されました）。追試によって，同じような研究を積み重ねれば，再現性がどの程度あるのか，また，同等の研究結果を統合して，その効果を検討することができるメタ分析も可能になります。そのため，「良い追試は，リサーチ・デザインのきちんと練られていない独創的な研究よりも，分野としての発展に貢献する可能性

が高い」（竹内・水本，2012，p. 341）と言えるでしょう。特に，「LMS を利用した教育効果の測定」という内容を考えてみると，CALL（Computer-assisted Language Learning）や教育工学の分野で似たようなアプローチの先行研究があると考えられるため，追試やメタ分析を行うことによって，知見の蓄積に役立つであろうと思われます。

2. 量的アプローチで教育効果の測定を行うときの注意点

2.1. 構成概念，妥当性・信頼性，一次元性

量的アプローチで用いるテストや質問紙では，「英語力」や「動機づけ」など，理論上は想定できる特性ではあっても，実際には観測できない構成概念（construct）の測定を行います。そのため，教育効果の測定で用いる測定道具は妥当性（validity）と信頼性（reliability）を持ったものでなくてはなりません。

妥当性の定義とその検証方法は時代と共に変化していますが（詳細は『外国語教育研究ハンドブック』第 2 章参照），「測定しようとしているものを正しく測定している程度」という定義は常に同じで，構成概念の測定において最も重要な観点になります。新しくテストや質問紙を作る場合は，必ず妥当性の検討・検証が必要であり，既存の測定道具を使う場合にも，妥当性の検討・検証がその開発段階でされているものを使用すべきです。

信頼性は，「その測定を何度繰り返しても同じ人には同じ結果が得られるだろうという精度」のことを指します。体重計に何度乗っても同じ結果が得られるように，テストや質問紙の得点も，同じ人が回答した結果がほぼ同じであれば「信頼性が高い」と考えられます。信頼性を示す指標（係数）としては，テストや質問紙を 1 度実施することで算出が可能な，クロンバックのアルファ（Cronbach’s α）が現在では最もよく用いられます。

「構成概念を妥当性・信頼性を持った道具を用いて測定を行う」という場合に，「測定しているものは 1 つの構成概念だけである」という一次元性（unidimensionality）の定義も覚

えておきましょう。テストや質問紙では，測定しようとするもの（構成概念）があって，それに対して数問の項目を用意し，その回答の合計得点を算出することで，例えば「この30問のリスニング・テストの受験者40名の平均点は15.51点（標準偏差3.21）で，この学習者の得点は19.82点であるので平均点より高い。」という判断や，「動機づけの質問紙のうち，外発的動機づけの尺度（5項目）で得点が高かったため，この学習者は外発的動機づけの高い学習者である。」などという判断を行います。このような形で合計得点を「素点」(raw score)で算出する方法を古典的テスト理論(classical test theory)と呼ばれることもあります。これは，項目応答理論（item response theory: IRT）という，近年，テストや質問紙の開発で用いられることが多い，測定の観点からはより好ましい方法があって，それと対比をするためにしばしば用いられる用語です（『外国語教育研究ハンドブック』第15章参照）。ただし，現在でもテストや質問紙を使った研究では，古典的テスト理論に基づく分析を行っていることがほとんどですので，言語能力をテストで測定する場合や，人の特性を質問紙で測定する場合には，「想定している構成概念1つだけを測定していて，素点を足し合わせた合計得点が，その測定しようとしている能力や特性を代表する指標となっているか」という，一次元性を常に考慮しなければなりません。

「合計得点に意味がある」(一次元性がある）ということは，測定において当たり前の事だと思うかもしれませんが，意外とそうではないこともあります。実際に項目分析をしてみると，その項目と合計得点の相関である項目-全体得点相関係数（item-total correlation coefficient）が低い項目もあります。これは，測定しようとしている構成概念以外のものを，その項目が測定している可能性を示唆しています。そのような場合は，該当項目を削除してから合計得点を計算するほうが，測定しようとしている構成概念をより的確に反映した得点になり，信頼性係数も高くなります。テストの合計得点を絶対的なものと考えている人には，「ある項目を外してから合計得点を再度計算する」というような方法は違和感があるかもしれませんが，そちらのほうが測定の観点からは好ましいと言えます。

2.2. 教育効果の測定にふさわしい研究デザインと分析方法

教育（介入）効果の測定を行う場合には，指導を行う処置群（treatment group）と指導を行わない対照群（contrast group）とに「無作為に」振り分けるのが最も望ましい方法ですが（処置群は実験群，対照群は統制群とも呼ばれることがあります），実際には，教室環境で指導の効果を見る事が多いため，そのような統制はできずに，無作為に振り分けられていない通常のクラスのようなグループ（intact group）を対象とした，準実験デザイン（quasi-experimental design）となることが多いでしょう（南風原, 2001; 村井，2012）。その場合にも，ある教育介入の効果を検証するという目的であるのなら，処置群だけではなく対照群を設ける必要があります。また，教育介入後に行う事後テスト（posttest）のみで処置群と対照群を比較した場合には，もともとの能力の違いが結果に影響しているという可能性を排除できないため，教育介入前に行う事前テスト（pretest）を実施するべきです（テストは項目応答理論を使って同一の尺度に等化した別フォームを使用するのが望ましいですが，同じものを使う場合もあります。その場合，事前テストの内容を覚えていて，事後テストで影響が出るという練習効果に注意すべきです）。以下の表 1 にこれらの点をまとめました。教室環境で指導の効果を見る場合には，様々な制約があるため，ケース 2 やケース 3 のような研究デザインになる可能性もありますが，厳密な教育効果の検証では，できる限りケース 4 のような研究デザインが求められます。

表 1　教育効果の測定における準実験デザインの 4 ケース

ケース	グループ	測定	問題点	評価
1	処置群のみ	事後テストのみ	何も主張できない	×
2	処置群のみ	事前・事後テスト	指導による効果かどうかわからない	△
3	処置群・対照群	事後テストのみ	もともとの学力の差が影響しうる	×
4	処置群・対照群	事前・事後テスト	無作為な割り当てができていない	○

注　グループが 2 群以上で事後テストのあとにさらに遅延テスト（delayed test）がある場合もある。

ただし，教室環境で対照群を設定するのは倫理的にも問題がありますので，先行研究で

教育効果があるとわかっているような指導内容であれば，ケース 2 のように処置群のみに事前・事後テストを行うことも考えられます。その場合には，先行研究でどの程度の効果があるとわかっているかを詳しく説明すべきです。また，対照群を設けた場合には，実験期間後には処置群と同じ指導を行うなど，倫理的に問題がないように配慮すべきです。

このように処置群と対照群を作り，指導の前と後に事前・事後テストを実施する準実験デザイン（2 群事前事後テストデザイン）では，以下の表 2 のようなデータが得られます。表 3 は表 2 のデータを集計した記述統計です。また図 3 はこの結果を図示したものになります。

表 2　2 群事前事後テストデザインのデータ（仮想）

名前 （ランダム生成）	グループ	事前テスト	事後テスト	差得点（事後 − 事前）
丸川　諒一		31	48	17
厚地　悠二		39	51	12
盛下　未季	処置群	56	67	11
江上　れな		47	44	-3
熊倉　実夢		29	33	4
（以下 25 名略）				
君塚　晴夏		36	42	6
奥芝　未菜		39	41	2
綾戸　海輝	対照群	39	44	5
宮司　恭祐		42	30	-12
伊島　雄樹		17	13	-4
（以下 25 名略）				

表 3　データの記述統計

グループ	人数	テスト	平均	標準偏差
処置群	30	事前	37.73	9.45
		事後	49.57	11.01
対照群	30	事前	38.37	10.58
		事後	40.13	12.03

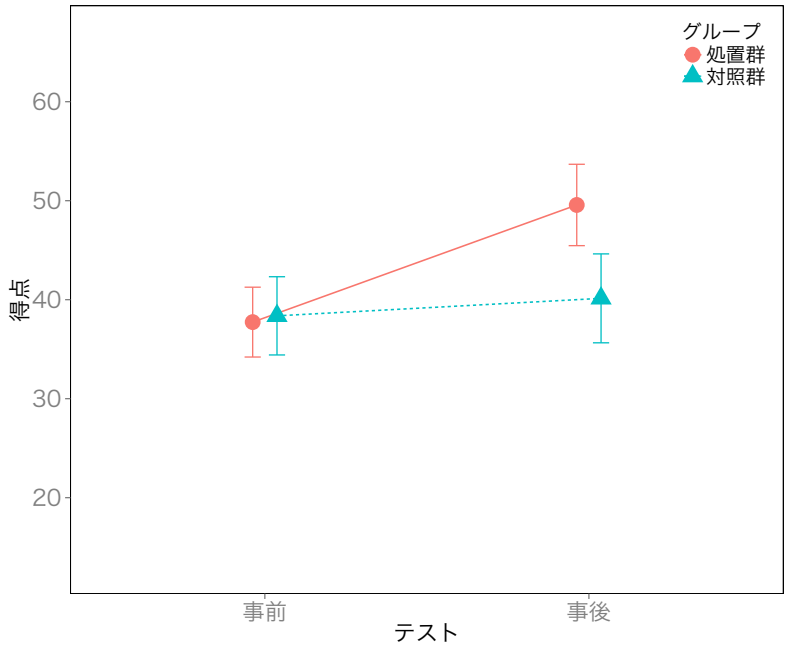

図 3.　結果の図示（エラー・バーは 95%信頼区間を示す）

このようなデータの分析方法として,『外国語教育研究ハンドブック』では二元配置分散分析（two-way ANOVA）を行い，グループ（処置群・対照群）とテスト（事前・事後）の交互作用（interaction）に注目するという方法を紹介しています。この二元配置分散分析での交互作用の結果と，差得点（事後テストから事前テストの得点を引いたもの）を使った群間の *t* 検定は同じ結果が得られます。ただし，差得点（gain score / difference score）は信頼性が低い場合も往々にしてあるため, 差得点の信頼性係数（Williams & Zimmerman, 1996）を確認して値が低い場合には，相関係数などで他の変数との関係を探る目的で使用しないようにしましょう（Dimitrov & Rumrill, 2003）。これらの方法よりもふさわしい方法として，事前テストの得点を共変量（covariate）として，事後テストの得点を（従属変数として）使い，共分散分析（analysis of covariance: ANCOVA）を行うという方法が，精度が高く推奨されます（Dörnyei, 2007; 南風原, 2001; 吉田，2006）。ただし，事前テストで 2 群の得点差が大きい場合は，事前テストでもともと得点が高い群が，「変化が大きい」という結果になってしまい，誤った解釈につながるという指摘もありますので（Cribbie & Jamieson, 2004），

図 3 のように結果を図示することで解釈に間違いがないか確認するようにしましょう。

事前・事後のテストデザインで教育効果の測定を行う際に，必ず知っておかなければならないことがあります。それは「平均への回帰」（regression to the mean）と呼ばれる現象です。これは，教育介入の事前・事後でデータを取った場合，事前テストで平均点よりも高い得点だった人は事後テストで得点が低くなり，事前テストで平均点よりも低い得点だった人は事後テストで点数が高くなる傾向があるというものです。聞いたことがなければ，にわかには信じ難い話ですが，平均への回帰は普遍的な現象であるため，「上位群が事前テストと比べて事後テストで得点が下がった。」「下位群が事前テストと比べて事後テストで得点が上がった。」というような結果が得られた場合には，もしかすると，教育介入の効果ではなく，平均への回帰による変化である可能性もあるので，解釈に注意が必要です（上位群と下位群に分けて分析を行った場合も同じくです)。印南（2012）は，このような平均への回帰への対処として，(a) データ収集前の対策（[a1] 2 群事前事後テストデザインを用いる，[a2] 共分散となり得る第三の要因を測定しておく，[a3] 信頼性の高いテストを使用する）と，(b) データ収集後の対策（[b1] 視覚化，[b2] 差得点と事前テスト得点の相関がマイナスでないかの確認，[b3] 実際の得点の変化が回帰効果を上回っているかの確認，[b4] 共分散分析や共分散構造分析などの統計的手法を用い回帰効果の影響を調整する）などの方法を紹介しています。

また，(多くの研究がそうであるように）事前・事後テストの得点を素点による合計得点で算出する場合は，古典的テスト理論の欠点である，「そのテストの受験者の能力とテスト項目の難易度によって，結果と点数の持つ意味が変わる」という大きな問題があります。特に，事前テストでほとんどの受験者が満点に近い得点であった場合には，天井効果（ceiling effect）が起こっていて，実際にはもっと高い能力があるにもかかわらず，満点以上の点数（難易度）がないために，それ以上の能力の正確な測定ができていないと考えられます。床面効果（floor effect）は，天井効果の逆で 0 点以下の能力であっても，0 点が取りうる最低の得点であるため，それ以下の能力の正確な測定ができていないケースです。そのよう

な問題には，前述の項目応答理論を用いることで対処することが可能な場合もあります（Dimitrov & Rumrill, 2003）。

上記まででは，教育介入の事前・事後の 2 回のみの測定を行った前提で説明をしてきましたが，個人の能力における変化の軌跡（trajectory）やプロセスを正しく測定するには，3 回以上測定を繰り返すほうが好ましいと言われています。その理由として，Singer and Willet（2003）は，「2 回の測定では真の変化と測定誤差を区別することができない」（p. 10）ということを指摘しています。これまでは，このような 3 回以上繰り返し測定を行ったデータに対しては，繰り返しあり（反復測定）の分散分析（repeated measures ANOVA）を用いることが多かったのですが（『外国語教育研究ハンドブック』参照），最近では，データの構造や個人の変化の測定により適切な方法である，マルチレベルモデル（multilevel model）や，潜在曲線モデル（latent growth curve modeling）の使用が増えてきています。マルチレベルモデルは，分野によって，線形混合モデル（linear mixed model），階層線形モデル（hierarchical linear model: HLM）とも呼ばれます（詳細は，水本，2012 を参照）。潜在曲線モデルは，構造方程式モデリング（structural equation modeling: SEM）を用いたものです（小杉・清水，2013）。マルチレベルモデルと潜在曲線モデルの違いは，徳岡（2013, p. 207）や Byrne, Lam, and Fielding （2008）などに詳しい説明がありますが，どちらも縦断データを分析する適切な方法で，推定方法に違いはあるものの似ている結果が得られます。

教育効果の測定結果を報告する際には，追試ができるような情報を書くことを心がけましょう。具体的には，（a）サンプルサイズ（人数），（b）平均値，（c）標準偏差，（d）（項目間や尺度間の）相関係数，（e）測定道具の信頼性係数などが記載してあれば，結果の再現を行うことができます。また，これらの情報に加えて，結果はわかりやすく図示するように心がけましょう（水本，2014）。結果の解釈においては，効果量（effect size）が実質的な差を確認するのに役立ちます（水本・竹内, 2008）。特に，これまでの同じような研究結果を統合して比べる，メタ分析で得られている効果量を，教育介入の実質的効果のベンチマークとして使用することも可能です。例えば，コンピュータを使って外国語の指導を行った

研究を対象とした Grgurović, Chapelle, and Shelley（2013）のメタ分析では，事前・事後テストのデザインで，効果量（Hedges' *g*）が 0.35（95%信頼区間 [0.26, 0.44]）と報告しているので，この値を越えるものはある程度の効果があると考える事ができます。

2.3. テストの使用目的と分析方法の一致

「教育効果の測定」という目的に関連して，最後にもう 1 つだけ覚えておいてほしいことがあります。それは，テストの使用目的と分析方法を一致させるということです。教育効果の測定を行うときに，前節までの考え方として指導の内容に直接は関連のない習熟度（proficiency）を測定するテストを用いることが多いと思います。英語のテストで言うと，TOEIC や TOEFL がこのようなテストですが，この種のテストは集団基準準拠テスト（norm-referenced test: NRT）と呼ばれ, テスト受験者の能力を他の受験者と相対的に比較し，入学試験などで合格・不合格を選別するような目的で使用されます。そのような目的のため，集団基準準拠テストの問題は，通常，ある一定期間の指導で行おうとしている内容以外のものを多く含みます。つまり，教育効果の測定に集団基準テストを用いても，習熟度に変化が現れるには時間がかかるため，指導によって何らかの効果があったにもかかわらず，それが得点に反映されないということも十分あり得ます。例えば，一般的にテスト対策は得点向上にある程度の効果が得られるということがわかっていますが（Farnsworth, 2013），TOEIC では信頼できる点数向上のためには，50 時間程度の集中的な指導が必要であるという結果も報告されており（Yamada & Ross, 2006），大学の授業で週 1 回 90 分の授業を 15 回行ったとしても，合計 22.5 時間しか指導時間はないため，LMS を上手く活用すれば変化は現れるかもしれませんが，なかなか難しいということは想像できるでしょう。

一方，授業で指導した内容に基づいた項目しかテストに含まず，学習者がある特定の知識やスキルがあるかどうか，もしくは特定のレベルに到達しているかを診断する目的のテストを，目標規準準拠テスト（criterion-referenced test: CRT）と言います。また最近では，standards-based assessment と呼ばれることもあります（Geisinger, 2012）。目標規準準拠テス

トを使用するような状況では，指導もそのテストに関連したものを行い，カリキュラム（指導）とテスト（測定）がセットになっていることが理想的です。つまり，普段の教室での指導の効果測定を「教育効果の測定」と考えるのであれば，集団基準準拠テスト（NRT）よりも，目標規準準拠テスト（CRT）を積極的に使用していくべきだということがわかるでしょう。そもそもの目的が違うことからも，目標規準準拠テスト（CRT）では，妥当性・信頼性，項目分析などの考え方が，集団基準準拠テスト（NRT）とは違いますが（Brown, 2005; Geisinger, 2012; Hudson & Brown, 2002; Kumazawa, 2007; 大友・中村・小泉，2009），指導やカリキュラムを改善したり，学習者がどのような内容を習得できていないかという点を検討する上で，非常に有益な情報が得られます。教育効果の測定においても，その目的に応じて，テストの使用目的（集団基準準拠テスト・目標規準準拠テスト）と分析方法を一致させることが重要であるということを覚えておきましょう。

謝辞

本稿の執筆にあたり，小泉利恵先生（順天堂大学），印南 洋先生（芝浦工業大学），熊澤孝昭先生（関東学院大学）に建設的なコメントとフィードバックをいただきました。ここに記して感謝いたします。

参考文献

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment* (New ed.). New York: McGraw-Hill.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge

University Press.

Byrne, B. M., Lam, W. W., & Fielding, R. (2008). Measuring patterns of change in personality assessments: An annotated application of latent growth curve modeling. *Journal of personality assessment*, *90*, 536–546. doi:10.1080/00223890802388350

Cribbie, R. A., & Jamieson, J. (2004). Decreases in posttest variance and the measurement of change. *Methods of Psychological Research Online*, *9*, 37–55. Retrieved from http://www.dgps.de/fachgruppen/methoden/mpr-online/issue22/mpr124_10.pdf

Dimitrov, D. M., & Rumrill, Jr, P. D. (2003). Pretest-posttest designs and measurement of change. *Work: A Journal of Prevention, Assessment and Rehabilitation*, *20*, 159–165. Retrieved from http://iospress.metapress.com/content/7x9hgpq885t2yttq/

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford: Oxford University Press.

Farnsworth, T. (2013). Effects of targeted test preparation on scores of two tests of oral English as a Second Language. *TESOL Quarterly, 47,* 148–155. doi: 10.1002/tesq.75

Geisinger, K. F. (2012). Norm- and criterion-referenced testing. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology: Vol. 1: Foundations, planning, measures, and psychometrics* (pp. 371–393). Washington, D.C.: American Psychological Association.

Grgurović, M., Chapelle, C. A., & Shelley, M. C. (2013). A meta-analysis of effectiveness studies on computer technology-supported language learning. *ReCALL*, *25*, 165–198. doi:10.1017/S0958344013000013

南風原朝和 (2001).「準実験と単一事例実験」南風原朝和・市川伸一・下山晴彦（編著）『心理学研究法入門一調査・実験から実践まで』(pp. 123–152). 東京:東京大学出版会

印南 洋 (2012).「テスト得点解釈の留意点」卯城祐司（編著）『英語リーディングテストの考え方と作り方』(pp. 78–87). 東京:研究社

磯田貴道 (2007). 『授業への反応を通して捉える英語学習者の動機づけ』 早稲田大学大学院教育学研究科博士学位論文 Retrieved from https://dspace.wul.waseda.ac.jp/dspace/bitstream/2065/28785/3/Honbun-4635.pdf

小杉考司・清水裕士（編著）(2013). 『M-plus と R による構造方程式モデリング入門』京都: 北大路書房

Kumazawa, T. (2007). Criterion-referenced test administration designs and analyses. *Second Language Acquisition - Theory and Pedagogy: Proceedings of the 6th Annual JALT Pan-SIG Conference*. 65–74. Retrieved from https://jalt.org/pansig/2007/HTML/Kumazawa.htm

前田啓朗・山森光陽（編著）(2004).『英語教師のための教育データ分析入門―授業が変わるテスト・評価・研究』東京: 大修館書店

水本 篤 (2012).「階層線形モデル/マルチレベルモデル/線形混合モデル」Retrieved from http://mizumot.com/lablog/archives/179

水本 篤 (2014). 「量的データの分析・報告で気をつけたいこと」外国語教育メディア学会（LET）中部支部 外国語教育研究基礎研究部会 第 1 回年次例会 講演資料 Retrieved from http://www.slideshare.net/AtsushiMizumoto/let20140222

水本 篤・竹内 理 (2008).「研究論文における効果量の報告のために―基礎的概念と注意点―」『関西英語教育学会紀要 英語教育研究』*31*, 57–66. Retrieved from http://www.mizumot.com/files/EffectSize_KELES31.pdf

Mizumoto, A., Urano, K., & Maeda, H. (2014). A Systematic review of published articles in ARELE 1–24: Focusing on their themes, methods, and outcomes. *Annual Review of English Language Education in Japan, 25*, 33–48.

村井潤一郎 (2012). 「実験法」 村井潤一郎（編著）『Progress & Application 心理学研究法』(pp. 15–48). 東京: サイエンス社

大友賢二・中村洋一・小泉利恵（編著）(2009).『言語テスト:目標の到達と未到達』東京:NPO 法人 英語運用能力評価協会（ELPA）

西條剛央 (2005).『構造構成主義とは何か―次世代人間科学の原理』京都:北大路書房

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford: Oxford University Press.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, *83*, 758–765. Retrieved from http://beta.edtechpolicy.org/CourseInfo/edhd485/AssessmentCrisis.pdf

住 政二郎 (2012). 「質的研究入門―基盤概念を知るには」竹内理・水本篤(編著)『外国語教育研究ハンドブック―研究手法のより良い理解のために』(pp. 242–257). 東京:松柏社

髙木亜希子 (2014). 「質的研究の世界へようこそ」外国語教育メディア学会(LET)関西支部 メソドロジー研究部会 2013 年度第 3 回研究会 講演資料

竹内 理・水本 篤(編著) (2012).『外国語教育研究ハンドブック―研究手法のより良い理解のために』東京:松柏社

徳岡 大 (2013). 「潜在曲線モデル」小杉考司・清水裕士(編著)『M-plus と R による構造方程式モデリング入門』(pp. 188–207). 京都:北大路書房

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement*, *20*, 59–69. Retrieved from http://www.phys.lsu.edu/faculty/browne/MNS_Seminar/JournalArticles/Gain_Scores.pdf

Yamada, H., & Ross, S. (2006). *Meta-validation of institutional TOEIC research in Japan*. Paper presented at the 10th JLTA Annual Conference. Kyoto, Japan.

柳瀬陽介 (2006).「第二言語習得研究や英語教育研究の「立ち位置」について」 Retrieved from http://ha2.seikyou.ne.jp/home/yanase/essay06.html#060706

吉田寿夫 (2006). 「研究法についての学習と教育のあり方について思うこと,あれこれ」吉田寿夫(編著)『心理学研究法の新しいかたち』 (pp. 244–270). 東京:誠信書房