The Japan Association for
Language Education & Technology
Language Education & Technology 52（2015）105–131

日本の外国語教育研究における効果量・検定力・標本サイズ
— *Language Education & Technology* 掲載論文を対象にした事例分析 —

草薙 邦広
名古屋大学大学院生
日本学術振興会

水本 篤
関西大学

竹内 理
関西大学

Reviewing Effect Sizes, Statistical Powers, and Sample Sizes of Foreign Language Teaching Research in Japan: A Case of *Language Education & Technology*

KUSANAGI, Kunihiro
Graduate School, Nagoya University
The Japan Society for the Promotion of Science

MIZUMOTO, Atsushi
Kansai University

TAKEUCHI, Osamu
Kansai University

Abstract

The purpose of the present study was to examine the quality of statistical testing in foreign language teaching research in Japan. We reviewed *t*-tests, χ^2-tests and ANOVAs reported in the articles published in *Language Education & Technology* (LET) from vol. 38 to vol. 49, and calculated the post-hoc statistical power of each test. The findings of the present study were summarized as follows: (a) the sample size of most of the studies in LET ranged from 20 to 60, (b) the median of the effect sizes (in

the case of *t*-tests) showed middle to large levels (d = 0.40–0.80), but (c) the statistical powers of many studies signified severely low levels (almost the 80% of the two-sample *t*-tests failed to show the statistical power greater than .80). The tendencies were quite likely to have originated chiefly from the inappropriate designs of the experiments or surveys, especially, mismatches between the targeted effect size and the actual sample size. We assert the importance of setting proper sample sizes based on a priori power analysis and precision analysis.

Keywords: statistical testing, statistical power, effect size, research quality, review

1. 背景

国内外を問わず，今日の外国語教育研究における実験，調査，そして実践報告の大多数が量的研究法をもちいている（e.g., Mizumoto, Urano & Maeda, 2014; Plonsky, 2013, 2014; Plonsky & Gass, 2011）。そのなかでも，近年は研究法の質的向上，とくにメタ分析（e.g., Norris & Ortega, 2006）や頑健統計（robust statistics; e.g., Larson-Hall, 2012; Larson-Hall & Herrington, 2010）といった統計手法の援用が重要視されている。しかしながら，心理統計学の観点からみた場合，これまでの外国語教育研究における統計的手法のあり方は満足できるものでないとも報告されている（e.g., Larson-Hall & Plonsky, 2015; Plonsly, 2013, 2014; Plonsky & Gass, 2011）。国内においても，前田（2000, 2004, 2010）が統計手法の選択や，論文での報告に不備がみられるということを繰り返し指摘している。

本研究は，そのような量的研究法のなかでも，統計的検定についてあつかうものである。従来からさまざまな研究分野において，統計的検定がもつ問題点や限界点について繰り返し指摘されている（e.g., Kline, 2004; Norris, 2015）。とりわけ，(a) 帰無仮説を採択することができない，(b) 有意水準の設定に恣意性がのこる，(c) 調査者が任意にきめられる標本サイズ（n）に対して検定統計量やp値が依存する，といった問題点があるといわれている（大久保・岡田, 2012）。

2000年代以降の心理学分野においては，過度に統計的検定をもちいることについて疑義を呈し，効果量（effect size）の算出，検定力分析（power analysis）や各統計量における信頼区間（confidence interval）の推定について，その有効性を主張する動きがある。これは統計改革とよばれている（e.g., 大久保, 2009; 大久保・岡田, 2012）。実際に，American Psychological Association（APA，アメリカ心理学会）の論文出版マニュアル（American Psychological Association, 2009）では，論文中において効果量を記載することが明示的に義務づけられている。また，検定結果の解釈においてその検定力を考慮する必要性があることも主張されている。しかしながら，国内の研究論文ではそのような統計改革の兆候がみられず（大久保, 2009），効果量の記載や検定力分析の重要性は比較的軽視されてきた，という主張もある（e.g., 大久保・岡田, 2012; 鈴川・豊田, 2012）。

多くの場合，外国語教育研究は，手法上心理学研究に準じ，APAによる論文出版マニュアル（American Psychological Association, 2009）を論文執筆の要領とする場合が多いが，水本・竹内（2008）は，すくなくとも2008年度の段階では，国内の学会誌で効果量を報告している論文は非常にすくなく，効果量の考え方がひろく浸透しているとはいいがたい，と述べている。しかしながら，当該の分野において効果量の平易な解説資料の提供，そして計算ツールの無償公開をした水本・竹内（2008）以降は，外国語教育に関する統計関係の解説書（e.g., 平井, 2012; 竹内・水本, 2014）にも効果量についての解説が散見されるようになっている。このようなこともあってか，近年では研究者の間にも効果量の概念が比較的浸透してきているようにみえる。

一方で，検定力分析の使用については，水本・竹内（2011）といった概説的資料が見られるものの，我が国の外国語教育研究分野において，効果量ほどの浸透がみられるかについては疑問がのこる。効果量が研究の対象となる効果の大きさを示すとするならば，検定力とは，「平均差が0である」というような帰無仮説が真実として偽である場合に，当該の検定が帰無仮説を棄却する確率である（詳しくは後述）。すなわち，統計的に有意な差が真として存在すると仮定したとき，検定が「事実を正しく評価する度合い」であるのだから，検定力を考慮しない検定結果の解釈は，実質科学的な知見を大きく歪める可能性すらある。また，検定力の低い検定の場合，同条件での実験が当初の結果を再現する確率は低い（e.g., Cohen, 1994）。

近年，特定の研究分野における研究論文の検定力を事例的に分析する試みが数多くなされるようになってきている（国内における心理学分野の例として，杉澤, 1999; 鈴川・豊田, 2011, 2012）。おどろくべきことに，このような研究では，「当該分野の半数程度の検定結果が望ましいとされる検定力を下回る」と報告されている。しかしながら，これまでの我が国の外国語教育研究における検定力の実態に焦点をあてた事例的分析は非常に限られている。

そこで本稿では，効果量および検定力について概観し，外国語教育メディア学会発行の機関誌である *Language Education & Technology*（LET）における2001–2012年号掲載論文計343編を対象とした検定力の分析結果を報告する。対象となった主な検定は，外国語教育研究分野において最も代表的な分析手法である，平均差についての*t*検定である。対象とした論文のうち，これらの検定結果について，効果量および検定力を集計し，我が国の外国語教育研究における検定力はどの程度であるかについて，その実態をあきらかにすることを第一の目的とした。次節は，効果量と検定力分析についての手法的概観と，これまでの研究論文における効果量や検定力の事例調査をおこなった先行研究の紹介にあてている。第三節と第四節は，LET掲載論文を対象とした事例分析にあてている。さらに第五節では，効果量，検定力分析および信頼区間の重要性を主張し，実験・調査計画における標本サイズの決定手順などについて提言をおこないたい。

2. 効果量と検定力

2.1 統計的検定と効果量の関係

効果量は一般的に，実験（調査）における効果の大きさ[1]を示す統計的な指標とされ

ている。ここでは，外国語教育研究に代表的だと考えられる，2 群のテストにおける平均値の比較を例として，効果量について概観する。A 群と B 群の平均値を比較した時，その平均差が 10 であったとする。平均差のみによる結果の解釈では，各群内のばらつきが考慮されていない。また，測定のスケール（点数など）に値が依存してしまい，解釈が限定的になる可能性もある。たとえば 10 点の平均差がどれくらいの学習効果の大きさを意味するのか，また別のテストではどれくらいの伸びにあたるのかなどは，平均差の検討のみではわからない。そこで，2 群のばらつきを加味し，さらに，測定単位に依存しない値をもちいて議論をおこなう必要性が生じる。

効果量の一種で，*d* 族とよばれる種類の統計量のなかでも，もっとも代表的な Cohen's *d* は，2 群の平均値（平均差）にそれぞれの標準偏差を加味することによって算出される。また，*d* はその定義から標準化平均差（standardized mean difference）[2] ともよばれる（Olejnik & Algina, 2000）。*d* は正負の符号をとり，理論的に，その大きさは 0 から無限大である。*d* は測定のスケールに依存せず，どのような尺度であっても，仮にその値が 1.00 だとした場合，2 群の平均差は 1 標準偏差程度だと解釈できる。*d* の式の一例を（1）に示す。*m* はそれぞれの平均値，*S* は標準偏差をあらわす。ここで，効果量の算出の過程において，標本サイズが加味されていない点にも再度注意されたい。

$$d = \frac{m_1 - m_2}{S} \tag{1}$$

一方，平均差の検定にもちいる検定統計量の *t* 値は，2 群（または 2 変数）の平均値，標準偏差，そのうえに標本サイズによってもとまる。*t* 値は仮に 2 群の平均値と標準偏差を変えずに，標本サイズのみを増やしていくと単調に増える傾向にある。一方，*t* 値と自由度に対応する有意確率は *t* 値の増加にともない，単調に小さくなっていく。このことから，検定統計量とは「効果量と標本サイズの関数である」であるともいえる（e.g., 南風原, 1995, 2002; Rosenthal & Rosnow, 1984）。例として， *t* 検定における *t* 値の式を（2）に示す。

$$t = \frac{m_1 - m_2}{S} \times \sqrt{\frac{n_1 \times n_2}{n_1 + n_2}} \tag{2}$$

検定統計量のこのような特性から，*t* 値およびそれに対応する有意確率（*p*）は，しばしば研究者が決定する標本サイズの影響を過剰に受ける可能性がある。さらに，有意差が見出された論文の方が論文誌などで公刊される可能性が高いため，実質的な効果が小さいような現象も，研究者のコミュニティにおいて過大に評価されてしまう可能性がある。

したがって，統計的検定の有意性にのみに依存するのではなく，標本サイズが加味されない標準化平均差（効果量）も，検定統計量と同時に報告することがのぞましい（e.g., 水本・竹内, 2008; 大久保・岡田, 2012）。

効果量の指標には *d* 族のみならず *r* 族とよばれる相関係数に基づく種類のものもあり，また，*d* 族の効果量から *r* 族の効果量へ変換する式なども紹介されている（概説は，水本・竹内, 2008 など）。また，同じ *d* でも，2 群における標本サイズの差についての取り扱いや，繰り返しのあるデータにおける相関係数の考慮などについて，算出方法が多岐に渡る（e.g., 水本・竹内, 2011; 豊田, 2009）。そのため，具体的な算出方法については論文中で明示するとよいだろう（e.g., 大久保・岡田, 2012, p. 61）。

2.2 検定力分析

検定力（1 - β, power）とは，「帰無仮説が偽であるときに，検定が帰無仮説を棄却する確率」である。つまり，第二種の過誤（Type II Error）を犯さない確率である。また，対象となる効果が実際に存在するときに，検定がその効果を見逃さない程度ともいえる（e.g., Cohen, 1988）。検定力は計算上，有意水準，標本サイズ，および効果量に対応する関数ともとらえられる。そのため，この性質を利用して，上記4つの要素のうち，ほかの3要素がきまれば，任意の要素の値を逆算することができる（豊田, 2009）。例として，有意水準をα = .05とした場合の，対応のある*t*検定における，検定力，効果量，必要標本サイズ[3]の関係を図1として図示する。

図1からわかるように，対象とする効果量が大きければ大きいほど，そして目標とする検定力が大きければ大きいほど，必要標本サイズもそれにともなって大きくなる傾向がある。望ましい検定力の基準は，研究分野やその文脈に依存するが，.80をもって，十分であると慣習的にみなされる（Cohen, 1988）。これを一般にCohenの基準とよぶ。しかしこの基準は，決して固定的なものとしてとらえるべきではない。

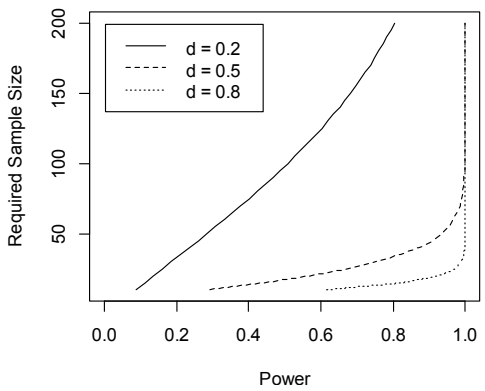

図 1. α = .05 における検定力・効果量・必要標本サイズの関係（対応のある *t* 検定の場合）

このような計算を，実験（調査）の事後的解釈，または研究活動における経済性の観点を踏まえた実験（調査）計画，とくに，標本サイズの決定のためにもちいる。これを検定力分析とよぶ。たとえば，豊田（2009）は検定力分析の主要な用法を以下の3つに分類している。

（a） 事前分析：任意の検定力，有意水準，効果量から必要な標本サイズをもとめる
（b） 事後分析：実際に得られた効果量と標本サイズ，任意の有意水準から検定力をもとめる
（c） 明日への分析：実際に得られた効果量と任意の（目標とする）検定力から必要な標本サイズをもとめる

これらの一連の分析は，十分な検定力を得ることを目標としてなされるものである。しかし，常に高い検定力がかならずしも望ましいわけではない（e.g., 南風原, 2002; 豊田, 2009）。とくに，研究者が唯一自由に決めることができる標本サイズは，検定力に対して単調に増加する関係にある。そのため，高すぎる検定力はその実験や調査の経済性を疑うものにもなりかねない。もちろん，低すぎる検定力は結果の信憑性に大きく関わるものであるのだから，問題視されるべきであることに変わりない。

また，効果量と検定力の関係も同様であり，標本サイズと有意水準を固定したときに，効果量が高くなればなるほど検定力も高くなる。そのため，小さい効果量を対象とする検定の検定力は必然的に低くなる傾向にあるため，検定力を望ましい水準（.80 など，e.g., Cohen, 1988）に保つためには，標本サイズを大きくする必要性がある。また，逆に大きな効果量を対象とする検定の標本サイズが小さくともよい可能性も大いにある。このように，研究者が関心を持つ効果量の大きさによって，十分な検定力をもって検定をおこなうことができる標本サイズは異なる。よって，研究者は興味のある現象に対して，あらかじめ見込まれる効果量の大きさに予測を立て，それにふさわしい標本サイズを決める必要がある（e.g., 南風原, 2002; 豊田, 2009）。

2.3 検定力の事例分析

上記のように検定力の重要性は認められるところであり，国内外を問わず，心理学研究の分野で，過去に公刊された研究の効果量および検定力を対象とした調査が複数おこなわれている（例：Cohen, 1962; Sedlmeier & Gigerenzer, 1989; Rossi, 1990; 杉澤, 1999; 鈴川・豊田, 2011, 2012）。1960年発行の*Journal of Abnormal and Social Psychology*掲載論文を対象にしたCohen（1962）は，70篇ほどの論文における標本サイズを調査した。Cohenは，これらの研究では，標本サイズの決定において検定力をもちいた事例はほとんどみられないと報告している。また，対象の論文における標本サイズから，任意の効果量（小・中・

大）[4]に対する検定力を計算したところ，中程度以下の効果量を検出できる研究は半分にも満たなかったと報告している。それからおよそ20年後の研究にあたるSedlmeier and Gigerenzer（1989）やRossi（1990）も同様の結果が見られたと報告している。

杉澤（1999）は，日本の『教育心理学研究』をおいて同種の分析をおこなっている。250編を対象とした分析でも，中程度以下の効果量を検出する標本サイズをもつ研究が半分以下であると報告している。鈴川・豊田（2011）は『認知科学』掲載論文を，鈴川・豊田（2012）は『心理学研究』掲載論文を対象に同様の分析をしており，これまでの事例分析研究と似通った結果を得ている。これらの研究では，一般的に効果量が低い検定においては大きい標本サイズをもつ研究が多いが，検定力分析をもちいて明示的に標本サイズを決定している研究は数少ないと主張している。

外国語教育研究においては，このような検定力の事例調査は非常に限られている。しかし，一般的な量的研究法の概観についてはいくつか散見される（e.g., Loewen & Gass, 2009; Norris, 2015; Plonsky, 2013, 2014; Plonsky & Gass, 2011; Plonsly & Oswald, 2014）。Plonsky and Gass（2011）とPlonsky（2013）は，当該分野のトップジャーナルである，*Language Learning*や*Studies in Second Langugae Acquisition*などの掲載論文の中でもちいられた研究手法や統計量の報告についてまとめている。Plonsky（2013）によれば，606本の対象論文のなかで，効果量を報告している研究は全体の26%（155件），信頼区間を報告している論文は5%（27件），そして検定力分析をおこなった検定は1%（6件）であったという。また，Loewen et al.（2014）は，応用言語学を専門とする研究者を対象として大規模な質問紙調査をおこない，研究者の統計に関する知識が不足していることを報告している。

Mizumoto et al.（2014）は*Annual Review of English Language Education in Japan*（ARELE）掲載論文を対象に，メタ分析的手法をもちいて総合的に国内における量的研究手法の質を調査している。Mizumoto et al.（2014）の結果によると，ARELE掲載論文における指導法効果研究では，平均的にみた場合，中程度の効果量を報告している研究が多かったものの（Hedge's $g = 0.76$ [0.59, 0.93]），63編中の27編（≒ 43%）が望ましい検定力の基準（.80）に満たなかった。そのうえ，およそ40%にあたる25編では検定力が.50以下だった。検定力が.50というのは，ほぼコイン・トスと同程度である（Mizumoto et al., 2014, p. 45）。

外国語教育研究では，なんらかの処遇の効果を研究対象とすることが多い。また，教育実践上のさまざまな問題から標本サイズを調整することは難しい（e.g., 草薙, 2014b）。さらに，完全無作為化による実験計画は組みにくく，倫理的配慮などから統制群をおけないケースもある。そのため，実験計画の効率化は分野特有の課題であるともいえる（e.g., 草薙, 2014b）。外国語教育研究のこのような状況が，上記のような低い検定力をまねいていると考えることは自然である。しかしながら，Plonsky（2013）が示しているように，検定力分析をもちいて標本サイズを決定した研究は，当該分野のトップジャーナルにおいて

もほぼ皆無である。心理学分野において南風原（1995）や杉澤（1999）が主張するように，外国語教育研究においても検定力を考慮した適切なデザインが強くもとめられるところである。

Mizumoto et al.（2014）のほかにも，平野（2011）や前田（2010）など，国内の外国語教育研究における研究手法や研究テーマの系統的レビューという先見性のある試みはいくつか見られるものの，これらの研究は統計的検定に関わるデザインの適切さに焦点を絞ったものではない。そのため，本研究では，日本の外国語教育研究における効果量，検定力，そして，とくに標本サイズに着目した事例分析をおこなう。

3. 調査方法

研究の対象となった論文は，外国語教育メディア学会の機関誌である LET の 2001 年度発行号（第 38 号）から 2012 年度発行号（第 49 号）の掲載論文であり，計 134 本である。[5] これらすべての論文を対象として，平均差の *t* 検定，χ^2 検定，および分散分析の統計手法と結果に関わる統計量をコーディングした。その中でも，LET 掲載論文においてもっとも代表的な統計手法である *t* 検定を中心として，効果量，検定力および標本サイズについて詳細に分析した。

3 つの統計手法に関するコーディングは以下のような手順でおこなった。なお，コーディングは著者および研究協力者 2 名（外国語教育研究および量的研究法を専門とする大学院生）が以下のコーディングスキームを参照した上でおこなった。

（a）平均差の *t* 検定

・対応のあるなし：論文中での記載を記録
・自由度（*df*）：論文中での記載を記録
・標本サイズ（*n*）：論文中での記載を記録
・効果量：Cohen's *d* を計算
　-自由度および標本サイズの記述がないものを除外
　-論文中の記述および *df* や *n* の整合性が取れないものを除外
　-記述統計量（*M* および *SD* の組）により *d* を計算
　-対応ありの場合，相関係数を考慮しない式によって *d* を計算 [4]
　-対応ありの場合で記述統計を報告していないケースを除外
　-対応なしの場合で *t* 値と *df* のみを報告しているケースは *t* 値から換算
　-その他，検定統計量の整合性が取れないものを除外
・検定力：効果量および標本サイズから計算
　-有意水準（α）を.05 として，得られた *d* と *n* から計算
　-論文とは無関係に両側の *t* 検定として計算

(b) χ^2 検定

＊事後の検定を含めず，適合度検定のみ[6]

・χ^2 値：論文中での記載を記録

・自由度（*df*）：論文中での記載を記録

・標本サイズ（*n*）：論文中での記載を記録

・効果量：φ およびクラメールの *V* を算出

(c) 分散分析

＊論文中での整合性が取れないものを除外

・第一自由度（df_1），第二自由度（df_2）：論文中での記載を記録

・標本サイズ（*n*）：論文中での記載を記録

・*F* 値：論文中での記載を記録

・効果量；η_p^2 を計算

-*F* 値と自由度から η_p^2 を算出

すべての検定のうち，統計量について整合性が取れないもの，ないし報告漏れがあるものはおよそ 20 件ほどであった。また，効果量について，すでに論文中に記載がある論文が 2 本あったが，著者が記述統計量から検算した。本研究では巻号年度によって論文を二種類に区分した（2001–2006，2007–2012）。また，Mizumoto et al.（2014）のように研究テーマによるカテゴリーを設けたところ，それぞれのカテゴリーに偏りができ，比較に十分でないと判断したため，研究テーマによる比較はおこなわなかった。

分析は，まず *t* 検定について，対象の効果量，標本サイズおよび検定力を記述統計を報告する。つぎに，標本サイズと効果量の関係について分析する。さらに，実際の標本サイズと得られた標本効果量に対して検定力.80 を満たす標本サイズについての関係性を散布図をもちいて検討する。

同様に χ^2 検定についての効果量，標本サイズおよび検定力の記述統計を報告した。また，分散分析は，デザイン（要因数，水準数，被験者内，被験者間，混合）によって効果量を比較することが困難であり，デザインごとに比較する場合，それぞれ十分なケースの数が得られなくなるため，分散分析の検定力については分析をおこなわなかった。

なお，効果量の計算には一般的な表計算ソフトを，検定力の計算には *G*Power*（Erdfelder, Faul & Buchner, 1996），および R の *pwr* パッケージ（Champery, 2012）をもちいた（解説は 豊田, 2009 など）。

4. 結果

4.1 平均差の*t*検定

最初に，*t* 検定についての分析結果を表 1 に示す。表 1 は，刊行時期（全体，2001–

2006 年度，2007–2012 年度）および対応のあるなしを区別して，*t* 検定の標本サイズ（*n*），効果量（*d*）および検定力の五数要約 [7] を示している。対象となった検定の数は 182 個であった。つぎに，刊行時期を区別せずに，それぞれの変数についての分布をヒストグラムで示している（図 2）。図 3 には同一のデータにおける累積分布を示している。

まず，標本サイズについて検討する。通常，対応のあるデータの方が有意差を得やすく，検定力も高くなるため，対応のある検定の方が標本サイズが小さくてよい場合が多い。しかしながら，本研究の対象論文においては，対応のあるなしの間で標本サイズがほぼ同等であった（図 2）。対応のある検定の場合の標本サイズは，最小値で 12 人，最大で 158 人，中央値で 43 人であった。また，対応のない検定では，最小値が 28 人，最大値で 150 人，中央値で 48 人であった。刊行時期を区別して中央値で比較してみると，後半の論文の方が全体的に標本サイズが大きくなる傾向がみられた。

表 1

対象論文における *t* 検定の標本サイズ（*n*），効果量（*d*）および検定力の五数要約

刊行時期	変数	対応	*k*	最小値	第一四分位点	中央値	第三四分位点	最大値
全体	標本サイズ	あり	133	12	16	43	53	158
		なし	49	28	30	48	74	150
	効果量	あり	133	0.00	0.38	0.85	1.27	3.60
		なし	49	0.00	0.22	0.46	0.79	5.47
	検定力	あり	133	.01	.49	.99	1.00	1.00
		なし	49	.01	.09	.38	.78	1.00
前半	標本サイズ	あり	62	12	19	38	45	158
		なし	33	30	30	40	54	97
	効果量	あり	62	0.00	0.27	0.82	1.76	3.60
		なし	33	0.00	0.14	0.44	0.79	5.47
	検定力	あり	62	.00	.24	.99	1.00	1.00
		なし	33	.05	.08	.23	.71	1.00
後半	標本サイズ	あり	71	16	43	52	66	140
		なし	16	28	48	73	90	150
	効果量	あり	71	0.03	0.58	0.86	1.16	2.40
		なし	16	0.07	0.34	0.55	0.80	1.20
	検定力	あり	71	.07	.98	1.00	1.00	1.00
		なし	16	.10	1.00	1.00	1.00	1.00

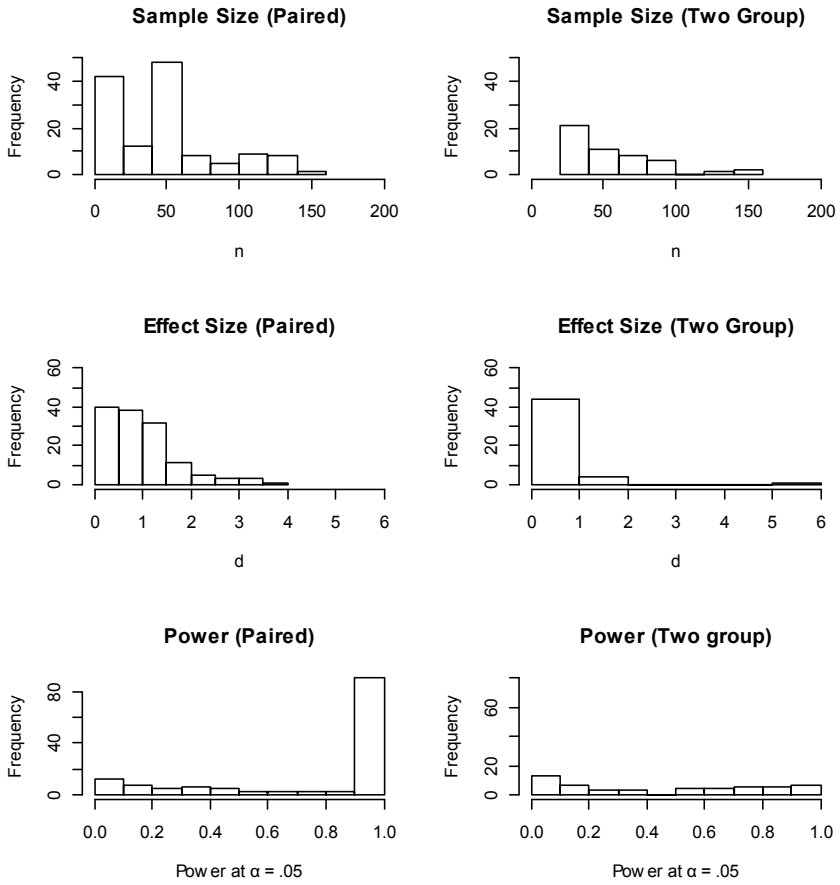

図 2. 全体を対象とした標本サイズ，効果量，検定力の分布を示すヒストグラム

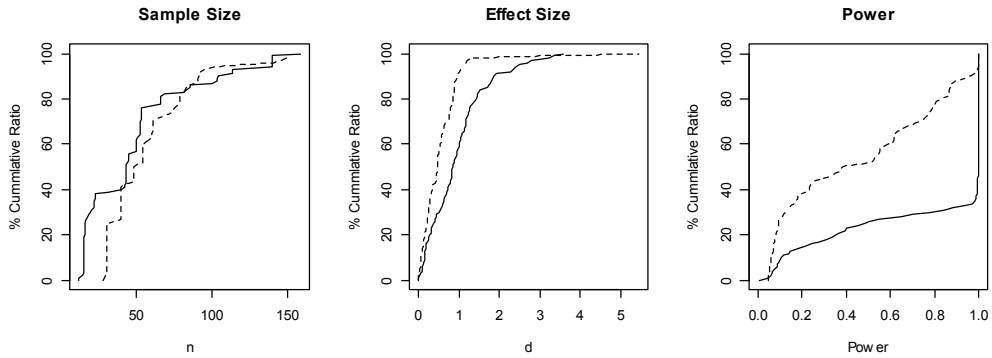

図 3. 全体を対象とした標本サイズ，効果量および検定力をあらわす経験累積分布関数。実線は対応のある *t* 検定，破線は対応のない *t* 検定をあらわす。

効果量について，対応のある検定の場合，最小値がおよそ 0，最大値で 3.60，中央値で 0.85 ほどであった。効果量の基準として，通常は 0.80 を大きな効果量とみなす（e.g., 水本・竹内, 2008）。このことから，半数以上の検定が効果量大を得ていると考えられる。これは Mizumoto et al.（2014）の結果とも一致する。一方，対応のない検定では，最小値がおよそ 0，最大値で 5.47，中央値は 0.46 であった。中央値である 0.46 は，およそ中程度の効果量とみなされる 0.40 よりもやや高い。対応のない検定の効果量は，対応のあるものよりも低いと考えられる。また，刊行時期を区別して比較したところ，大きな傾向の差はなかった。

つぎに，検定力について検討をおこなう。対応のある検定の検定力は，第一四分位点で.49，中央値以降はほぼ 1.00 を示した。このことから，およそ 70%程度のケースが最大レベルの検定力を示しているといえる。しかし，25%程度のケースは検定力が 0.50 以下でもある。対応のない検定の場合，第一四分位で.09，中央値で.38，第三四分位点では.78 だった。つまり，80%程度のケースは，望ましいとよばれる検定力の基準（.80）に満たない。また，半数以下のケースの検定力は 0.50 以下であった。

しかし，このような検定力の傾向は改善の兆しがみられる。比較するケースの数に偏りが見られるものの，後半の方が概して検定力が高い（表 1）。これは，効果量が前半と後半で同程度であったのに対して，標本サイズが大きくなったことがその原因であると考えられる。また，対応のある検定の方が検定力が高いのは，通常，対応のある検定の方が高い効果量を得やすい傾向にあり，さらに標本サイズが対応のあるなしで同程度であることからも理解できる。

また，対象とした検定の標本サイズに対して，それぞれ効果量が小・中・大程度であると仮定したときに得られる検定力について計算をおこなった（表 2）。

表 2
対象研究の標本サイズと効果量小・中・大にそれぞれ対応する検定力の五数要約

対応	*k*	*d*	検定力の五数要約				
			最小値	第一四分位点	中央値	第三四分位点	最大値
あり	133	0.20	.17	.18	.27	.40	.68
		0.50	.72	.75	.92	.99	1.00
		0.80	.98	.99	1.00	1.00	1.00
なし	49	0.20	.10	.12	.25	.30	.70
		0.50	.35	.47	.89	.95	1.00
		0.80	.71	.85	1.00	1.00	1.00

注：効果量は小を 0.20，中を 0.50，大を 0.80 とした（水本・竹内, 2008）。

対応のある検定では，効果量を小としたときに得られる検定力の中央値は.27 であり，最大値では.68 であった。つまり，本研究の場合，効果量小を対象とする場合に適した標本サイズをもつ検定は全くみられなかった。しかし効果量を中とした場合，およそ 80%のケースが望ましい検定力を満たし，大の場合はほぼすべてのケースが最大の検定力を示すことがわかった。

一方，対応のない検定では，効果量を小とした場合の中央値は.25，最大値で.70 であり，対応のある検定と同様に効果量小に対して適したケースはみられなかった。また，効果量中程度の場合であっても，半数程度のケースが望ましい検定力を満たせないことがわかった。しかし，効果量大の場合では，75%程度のケースが望ましい検定力を示した。

つぎに，標本サイズと効果量の関連性について検討する。図 4 に対応のあるなしによって区別した標本サイズと効果量についての散布図を示す。統計的検定の性質上，検定力を一定（たとえば.80）にするためには，効果量が小さければ標本サイズを大きくし，効果量が大きければ標本サイズを小さくしてもよい。効果量の大きさは研究者が対象とする現象に依存するが，標本サイズは研究者が決められるため，適正な実験（調査）計画にもとづく場合，複数の検定における効果量と標本サイズは関数を描くか，すくなくとも負の回帰係数を示すはずである。しかし，図 4 がそのような傾向を示すとはいいがたい。このことから，標本サイズの決定が対象とする現象の効果量の大きさとほぼ独立しておこなわれている可能性があることがわかる。

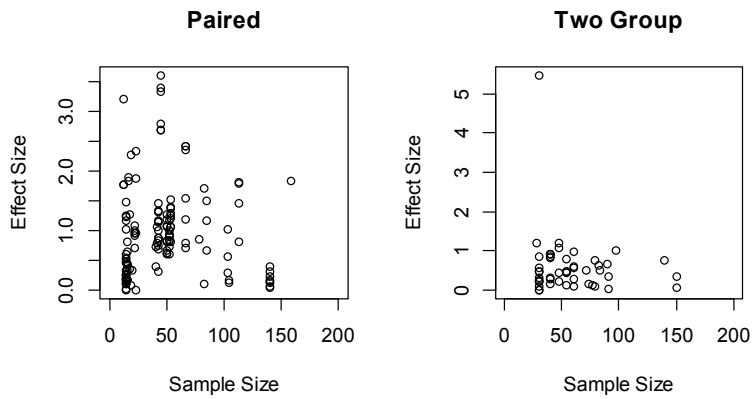

図 4. 標本サイズと効果量の関係を示す散布図

さらに明確に標本サイズの決定について検討するために，それぞれのケースの標本サイズと，それぞれのケースで得られた標本効果量および検定力.80 を満たす標本サイズ（事前の分析によって計算する。必要標本サイズとよぶ）を比較した。図 5 にその散布図を示す。また，計算上莫大な標本サイズに相当する場合があるため，便宜的に 500 以上の値をとるケース（38 件）は図から除外した。

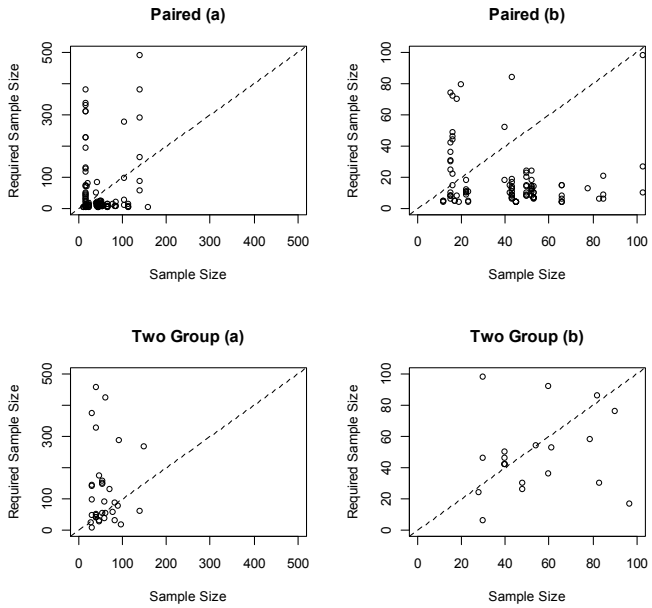

図 5. 実際の標本サイズおよび必要標本サイズの散布図。
（a）は 500 人まで，（b）は便宜的に 100 人までの範囲で示している。
破線は y = x を描いている。

標本サイズの設定が適正な場合，標本サイズと必要標本サイズは近似するか，すくなくとも線形モデルにはてはまるようなデータになるはずである。しかしながら，図 5 は明らかにそのような傾向を逸脱するものであった。また，必要標本サイズにかかわらず，とくに 20 人ほどの標本サイズに集中している傾向がみられる。さらに，図 5 の左側の図（a）からもわかるように，必要標本サイズとの大きな乖離を示す研究例も多い。

4.2 その他の検定：χ^2 検定と分散分析

ここからは χ^2 検定の分析結果について示す。χ^2 検定は全体で 6 本の論文，87 個の検定が対象となったが，同一の論文で繰り返し使用している例が多く，そのため，結果の一般化には十分に注意する必要がある。表 3 はその結果をあらわしている。

表 3
χ^2 検定の標本サイズ，効果量および検定力の五数要約

	最小値	第一四分位点	中央値	第三四分位点	最大値
標本サイズ	4	22	28	82	7121
効果量	.04	.11	.24	.35	.90
検定力	.05	.15	.33	.55	1.00

χ^2 検定の結果も同様に検定力が全体的に低く，90%程度のケースが望ましい検定力である.80 におよばない。*t* 検定に見られるように，一般に必要標本サイズに対して実際の標本サイズが小さいことが主な原因であると考えられる。

また，分散分析はデザインによって必要となる標本サイズが異なるため，効果量 η_p^2 のみを検討した（表 4）。しかし，η_p^2 は η^2 と異なり，実験計画の要因数に影響されないが，直接的な値の大きさの目安がないということに注意されたい。[8] なお，主効果および交互作用を含め，下位の分析については分析から除外してある。中央値である.09 は，誤差および当該の要因による平方和の和に対して，当該の要因による平方和が占める割合が 9% であることを示す。仮に 2 × 2 の交互作用について，この程度の効果量であれば，検定力.80 を満たす必要標本サイズは 90 ほどである。例えば、2 要因ともに対応なしのデザインの場合、1 群で約 23 人が必要ということである。しかし 2×2 のデザインで 90 人ほどの標本サイズをもつ研究は一般的にいって多くない。

表 4
分散分析の効果量についての五数要約

	最小値	第一四分位点	中央値	第三四分位点	最大値
効果量	< .01	.03	.09	.24	.94

注：$k = 306$.

4.3 まとめ

これらの結果をまとめると以下のようになる。

(a) 対象となった LET 掲載論文における標本サイズは，平均差の *t* 検定の場合，およそ 40 人程度が代表的であった。対応のあるなしで大きな差はなかった。標本サイズは刊行時期によってやや大きくなる傾向が見られた。

(b) 対象となった LET 掲載論文における統計的検定の効果量は全体的に中から大とよばれる基準に相当するものが多かった。これは Mizumoto et al., 2014 などの研究と同様の結果である。平均差の *t* 検定では，対応のある場合にその効果量が大きかった。また刊行時期による効果量の違いは大きくなかった。

(c) 対象となった LET 掲載論文における統計的検定の検定力は全体的に低い（これは他分野の多数の事例分析，そして Mizumoto et al., 2014 と同様の傾向である）。平均差の *t* 検定の場合，とくに対応のない場合にその傾向が顕著である。しかし刊行時期によって検定力はやや高くなった。χ^2 検定および分散分析の検定力も総じて同様に低いものと考えられる。

(d) 対象となった LET 掲載論文における標本サイズでは，効果量小程度に対して十分な検定力を満たすことができない研究がほとんどである。しかし中程度や大程度では十分な検定力を満たすことが多い。

(e) 対象となった LET 掲載論文における検定のデザインでは，標本サイズと効果量の適切な関係が見れられないものが多い。

(f) 対象となった LET 掲載論文における標本サイズと十分な検定力を満たす必要標本サイズの間には適切な関係が見られない。また，必要標本サイズとは関係なく，20 名前後のみの標本サイズを持つ研究が多い。

これらの結果は，研究者が統計的検定の仕組みや，適切な実験（調査）方法についての知識を持たないことに由来する可能性がある。たとえば，対応のあるなしでは，対応のある検定の方が必要標本サイズが小さくなる傾向があるが，対象論文では全体的に差がなかった。また，効果量が大きい場合は，必要標本サイズは小さくなるが，対象となった研究の標本サイズをみるとそのような傾向は見られない。さらに，一般的に研究分野の知見が集積すると，より小さな効果量について取りあつかう必要性が生じ，刊行時期などを区別すると平均的な効果量は低下し，標本サイズは増大すると予測できる（e.g., Plonsky & Oswald, 2014）。しかし，そのような傾向も見られなかった。これらのすべてが，すくなくとも「標本サイズの決定が不適切である」という点に帰着させることができるだろう。

これまでの研究でなされた統計的検定の検定力が低いということは，蓄積されてきた研究分野の知見をおびやかす可能性すらある。当該分野における統計的検定を，検定力の観点から十分なものにすることは，分野全体の喫緊の課題であるといえる。そこで，次節からはより望ましい実験（調査）のデザイン，とくに統計的検定の質を担保する標本サイズの決定方法について，いくつかの手法上の提言をおこないたい。

5. 望ましい統計的検定に向けて

ここでは，統計的検定をより望ましいものにするために外国語教育研究者がおこなうべき具体的な方法について提示する。一つめは，効果量による議論である。二つめは，信頼区間による議論である，三つめは事前の分析（検定力分析）をもちいた標本サイズの決定方法について，四つめは，誤差や信頼区間の幅による標本サイズの決定方法（精度分析）について，五つめはデータの可視化についてである。

5.1 効果量による議論

効果量は標本サイズとは独立した指標であるため，検定統計量に対応する *p* 値とは異なり，標本サイズの大きさに影響されない。そのため，統計的検定をもちいる際の問題点のひとつ（*p* 値の過度な標本サイズへの依存性）を避けることができる。効果量の計算に

は，さまざまな専用ツールが公開されており，計算式も複雑ではないために，容易にもとめることができる（実際の計算などに関しては，水本・竹内, 2008 など）。

しかし，効果量による議論ももちろん万能ではない。標本から得られた効果量は，標本効果量にすぎないのであり，標本効果量をさも母効果量それ自体であるかのように解釈してはならない。標本効果量はあくまでも母効果量の推定値にすぎないのであり，母効果量は基本的に不明である。また，標本効果量の推定精度，つまり誤差の大きさも標本サイズに依存する，ということを忘れてはならない。たとえば，統計的検定が標本サイズに依存するからといって，検定をおこなわずに，小標本によって推定された効果量をことさら重視して議論を進めることは，本末転倒である。

たとえば，平均差が 10，効果量（*d*）が 1.00 となるような正規分布乱数を 2 組セットで生成するシミュレーションをおこなうとする。標本サイズをそれぞれ，5，10，25，50，75，100，250，500，1,000 として 1,000 組の 2 標本をそれぞれ生成する。そしてそれらの標本の組において平均差と効果量をもとめ，散布図であらわすとする（図 6）。

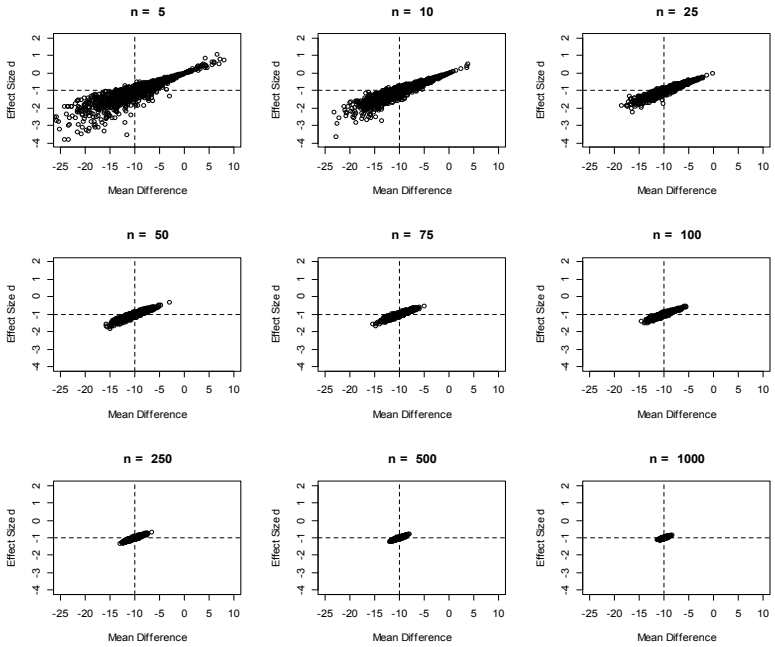

図 6. 平均差と効果量に関するシミュレーションの結果をあらわす散布図

乱数によるシミュレーションではあるが，ここでは母平均差が 10，母効果量が 1.00 であり，それぞれのケースを標本平均差および標本効果量とみなすことができる。誤差がない場合には、点線が交わった一点に値が集中するはずであるが、標本サイズが小さい場合，平均差と効果量が非常に大きな幅を取ることがわかる。また，平均差が小さくても効

果量は大きい，または効果量が小さいが平均差は大きい，といった標本も多くなる。しかし，標本サイズが大きくなるにつれ，その幅は小さくなっていき，平均差と効果量の関係も線形により近くなっていく。n = 1,000 になると，ほぼ点推定値に集約されていく。このように，効果量の推定精度自体（誤差の幅，信頼区間）が標本サイズの影響を受けている。よって，統計的検定の結果とは関係なく，効果量という指標をもちいると小標本でもよいなどということは決してありえない。

5.2 信頼区間による議論

信頼区間は，あらかじめ設定された確率をもって母数を区間として推定するものである。たとえば，母平均値について議論するのであれば，母平均の 95%（ときに 99%）信頼区間は[72.21, 83.32]であるというように，確率的下限値と上限値をもとめて提示する。ただし，信頼区間の厳密な解釈は複雑であり，「ある標本値に対して，その母数として確率論的に整合的である区間」といった程度の理解で実務的に問題ない（詳しくは，南風原, 2014; 大久保・岡田, 2012 など）。母平均値における 95%の信頼区間をもとめるには，以下のような式をもちいる。ここでの t は確率と自由度（n-1）に対応する t 値である。自由度が∞で 95%の場合，正規分布に近似するためおよそ 1.96 である。

$$95\%\ \mathrm{CI} = M \pm t_{critical} \times \frac{SD}{\sqrt{n}} \qquad (3)$$

信頼区間は効果量とは異なり，標本サイズからは独立でない。標本サイズが大きくなれば大きくなるほど信頼区間は狭くなる傾向にある。さらに，信頼区間は測定のスケールに依存するため，測定具が変われば単純な比較をすることができない。しかし測定のスケールにおける，実質科学的な差を議論することには適している。信頼区間は母平均のみならず，さまざまな統計量についてもとめることができる（大久保・岡田, 2012）。たとえば，平均差，相関係数，回帰係数などはもちろんのこと，効果量の信頼区間ももとめることができる（詳しくは後述）。また，母数の確率分布が未知の場合であっても，ブートストラップ法（bootstrapping）[9]をもちいることによって，理論的にはどのような指標であっても近似的な信頼区間の算出をすることができる。そのため，外国語教育研究分野においても，ブートストラップ法をもちいる信頼区間の計算が，近年研究者の注目を浴びている（e.g., 草薙, 2014a; Larson-Hall & Herrington, 2010; Plonsky, Egbert & Laflair, 2014）。また，ベイズ統計学の手法[10]のひとつであるベイズ信用区間（e.g., Edwards, Lindman & Savage, 1963）も，その重要性がしだいに主張されるようになってきている。

また，APA の論文出版マニュアル（American Psychological Association, 2009）も信頼区間の報告を強く推奨している。とくに，統計的検定をおこなう際には，平均値のみなら

ず，平均差や効果量の信頼区間も検定結果に付記するとよいだろう。[11]

5.3 検定力分析による標本サイズの決定

水本・竹内（2011）や豊田（2009）など，国内でも検定力分析に関する優れた概説書は見られる。しかし，検定力分析にもとづいた実験（調査）の設計は浸透していない。ここでは，実験（調査）の前に適切な標本サイズを決定する事前の検定力分析について，簡単に触れる。

事前の分析は，先述のとおり，任意の有意水準（通常，.05），実験の目安となる効果量（先行研究などと同様のレベルにするか，未知の場合は中程度など），そして目標とする検定力（通常.80）を満たす標本サイズを計算する方法である。たとえば，対応ありの t 検定で，同テーマの先行研究で得られた効果量が小（d = 0.20）であったとする。このときに，α = .05，d = 0.20，Power = .80 を満たす最小の n は，199 である。仮に d = 0.50 であれば，n = 34 となり，d = 0.80 であれば，n = 15 というようになる。

このような標本サイズの決め方をもちいる場合，論文の中でその概要を明示的に報告するべきである。たとえば，実験参加者に関わる論文の節のなかで，以下のように記述するとよいだろう。

> **本実験の標本サイズ（n = 18）は，実験実施前に検定力分析をもちいて決定した。本研究が対象とする a 指導法が b 能力におよぼす影響の大きさについて，本研究の主要な先行研究である○○○（2000）の実験の結果では，効果量大（d = 0.80）を示している。検定力分析の結果，d = 0.80，α = .05，Power = .80 を満たす最小標本サイズは 15 であることがわかった。本研究の標本サイズはこれをやや上回るものであるため，仮に先行研究と同等の効果が得られた場合，適切に有意差を検出することが期待される。**

検定力分析を実際におこなうためには，*G*Power* という専用のソフトウェア（Erdfelder, Faul & Buchner, 1996）をもちいるとよい。このソフトウェアは無償公開されており，操作も困難なものではない。また R の *pwr* パッケージについては豊田（2009）が詳細な解説をおこなっている。

5.4 誤差や信頼区間の幅による標本サイズの決定

信頼区間は標本サイズとの対応関係があるため，それを逆に利用して，任意の信頼区間の幅を満たす標本サイズを逆算することができる（e.g., 南風原, 2002; 永田, 2003）。これを精度分析（precision analysis）とよぶ。たとえば，母平均値の信頼区間を考えるとき，分散（または標準偏差）がわかっているときには，(4) 式が標準誤差（*SE*）の定義式であるのだから，(5) 式のように解けばよい。このようにして，比率，平均差，効果量，相

関係数などについて，任意の誤差を満たす標本サイズをもとめることができる。また，以下の式における標準誤差を信頼区間の幅とすると，任意の信頼区間の幅を満たす標本サイズも同様にもとめることができる。

$$SE = \frac{SD}{\sqrt{n}} \tag{4}$$

$$n = \left(\frac{SD}{SE}\right)^2 \tag{5}$$

また，効果量 d の場合における標準誤差は（6）式で計算できる（e.g., 南風原, 2002）。さらに（7）式は信頼区間の式である。これらの式を n について解くことで任意の効果量の 95%信頼区間を満たす標本サイズももとめることができる。

$$SE_d = \sqrt{\left(\frac{n_1 + n_2}{n_1 \times n_2}\right) + \left(\frac{d^2}{2(n_1 + n_2 - 2)}\right)} \tag{6}$$

$$95\%\ \mathrm{CI} = d \pm 1.96 \times SE_d \tag{7}$$

効果量にもとづいて議論をおこなうときは，この方法による標本サイズの決め方がとくに有効である。たとえば，効果量は通常 0.20，0.50，0.80 といった基準に関連づけられているため，そういった基準や 0 をまたがない信頼区間の幅を満たすような標本サイズに決めるとよい。検定力分析による標本サイズの決定と同様に，誤差や信頼区間の幅による標本サイズの決定についての情報も論文の中で明示的に記すべきである。以下にそのような報告の一例を示す。

本研究における標本サイズ（n = 60）は，効果量とその信頼区間の幅を考慮して決定した。本研究が対象とする a 指導法が b 能力におよぼす影響の大きさについて，本研究の主要な先行研究である○○○（2000）の実験の結果では，効果量大を示している。この効果量を d = 0.80，実質的な議論に十分であると思われる d の推定精度（95% CI）を± 0.40 とした場合，この幅を満たす最小の標本サイズはおよそ n = 50 であった。

残念ながら，現在，外国語教育研究においてこのように誤差や信頼区間の幅，とくに効果量の誤差や信頼区間の幅にもとづく標本サイズの決定がなされているとはいいがたい。

今後この方法が研究者の間に広まることが望まれる。

このような計算は表計算ソフトでも十分に可能である。任意の誤差や信頼区間を満たす標本サイズを半自動的に計算するスプレッドシート（*.xlsx*）を，第一著者のウェブサイトで公開している。[12]

5.5 可視化をもちいた方法

統計的検定の二値的な判断のみに頼るのではなく，標本におけるデータのばらつきや，母数の信頼区間などを総合的に検討することも，効果量や信頼区間による議論，および適正な標本サイズの設定と同程度に重要である。また，データのばらつきや信頼区間を適切に可視化することは，研究の質を高めるもっとも効率的な方法のひとつである。

データの可視化には，標本におけるデータについて豊富な情報を提示することを目的とする「記述用の可視化」と，主に複数の変数における母数や信頼区間などを同時に比較するための「解析用の可視化」がある（草薙, 2014c）。

「記述用の可視化」の方法としては，標本の五数要約をあらわす箱ひげ図，または全ケースの値を要約せずにそのままあらわす蜂群図などがある。このような可視化をおこなうことによって，慎重にデータを検討することができる。また，統計的検定をする必要があるか，検定の条件は満たされるか，などを確認することもできる。そもそも，外国語教育研究においては推測統計をおこなう必要がない研究テーマもあることも忘れてはならない。図 7 に箱ひげ図の例を示す。

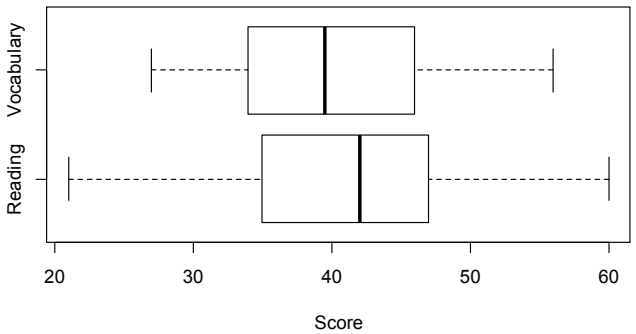

図 7. 標本の五数要約をあらわす箱ひげ図の例

「解析用の可視化」においては，信頼区間などを誤差棒であらわすことが効果的である。図 8 に信頼区間をつけた折れ線グラフの例を示す。グラフにおける誤差棒は，標準偏差，標準誤差，予測区間，信頼区間など，異なる範囲を示す場合があるため，どのような範囲をあらわしているか明示的に書く必要がある。[13]

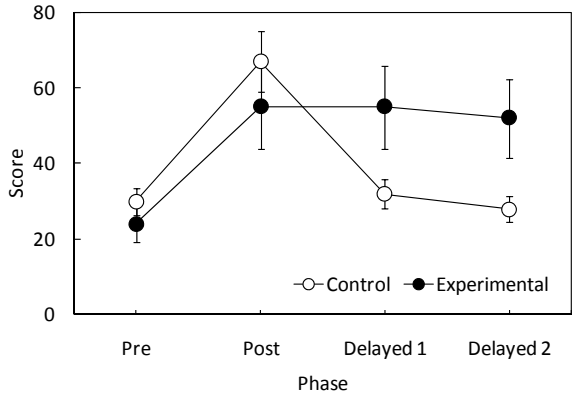

図 8. 母平均値の 95%信頼区間を付記した折れ線グラフの例

6. 結論

本研究では，10 年間の LET 掲載論文における統計的検定を分析することによって，日本の外国語教育研究において，(a) 検定力が低い研究が多い，(b) それは標本サイズの設定が不適切であることに由来する可能性がある，ということをあきらかにした。このような現状をよりよくするためには，(a) 効果量，(b) 信頼区間，(c) 検定力分析をもちいた標本サイズの決定，(d) 誤差や信頼区間にもとづく標本サイズの決定，(e) データの可視化，といったさまざまな統計手法上の理解が必要であると指摘し，これらの手法の概説を試みた。そのなかでも，とくに標本サイズの決定方法の重要性を主張した。

しかしながら，本研究には以下のように，いくつかの欠点がある。第一に，本研究の結果をもって，日本の外国語教育研究全般に一般化をおこなうことはできない。本研究は日本の外国語教育研究においては，代表的なジャーナルのひとつである LET を対象としている。しかし，単一のジャーナルを対象としてるため，テーマなどによって偏りがある可能性がある。今後，他誌における同様の事例研究と結果を照合する必要がある。

さらに，これまでの検定力における事例研究にくらべ，本研究は，対象研究数が多くない点も欠点としてあげられる。この点も一般化の可能性をより狭めるものである。また，*t* 検定のつぎに代表的な手法である分散分析については，デザインなどを詳細に分けて検討することができなかった。今後，対象とする論文をさらに増やすなどして，分散分析における検定力を調査することも必要である。

刊行時期，研究テーマ（学習者の能力，知識，心理特性，指導効果 etc.）や研究手法（質問紙調査，心理学的実験，テスト）などによっても，標本サイズ，効果量および検定力の実態は大きく異なると考えられる。それぞれの研究テーマや研究手法を専門とする研究者が，そのテーマや方法に関わる実質科学的な，そして専門的な知見との兼ね合いのな

かで統計的手法について議論することが強く望まれる。

結語となるが，研究をおこなううえで，統計的手法の洗練は非常に重要なことではある。しかし，そもそも，測定をおこなおうとするもの，それは得てして抽象的な概念である構成概念（construct）であったり，そしてそれは，日頃から自らが教育活動において関わる学生や生徒についてのことであったりもする。統計的に望ましい標本サイズを決めることは可能であり，もちろん望ましい標本サイズをもって研究に取り組むべきである。しかし，そのまえに，構成概念の実質科学的な側面についての理解，そして研究の対象となる現象の理解のために，それぞれの学生や生徒についての観察をけっしておろそかにしてはいけない。また，必要標本サイズが大きいからといって無闇に，（しばしば指導効果がより小さくなるような）統制群に多数の学生や生徒を機械的に割りあてたり，実質的な議論に必要となる以上に標本サイズを増やしたりすることは厳に慎まれるべきである。統計的に適切な標本サイズというものも，そもそも統計的手法の質を保証するひとつのものに過ぎないということを忘れてはならない。

注

1. 効果の大きさとは，実験操作がおよぼす影響の大きさや変数間の効果の強さをあらわす（e.g., 水本・竹内, 2008）。また，より抽象的に捉えると，あるデータのばらつきの中における，研究者が設定する任意のばらつきの大きさともいえる。一方，それ以外のばらつきを，通常は誤差ととらえる。
2. 効果量には，標準化した効果量（標準化平均差）と標準化しない効果量があり，後者を単純効果量（simple effect size）とよぶ（Frick, 1999）。単純効果量には平均差（m_1-m_2）などがある。標準化平均差は測定のスケールを失うため，測定のスケールが重要な場合は単純効果量を考慮するとよいとされる。
3. 検定力分析における必要標本サイズは，対応のある場合には，データの組の数をあらわし，対応のない場合は，群ごとの人数を足した総数をあらわしたりする場合があることに十分に注意されたい。
4. 連続量である効果量に対して，「小」「中」「大」といった語を対応づけることには注意が必要である。Cohen は，効果量の値の目安として，$d = 0.20$ 程度を「小」，0.50 程度を「中」，0.80 以上を「大」と分類し，この分類が慣習的にもっとも広く使用されているが，これになにかの数学的根拠があるわけではない（Cohen, 1988）。効果量の値は，個々の研究の文脈に依存するため，その値を総合的に検討することが望まれる。Plonsky and Oswald（2014）は，第二言語習得研究における既刊論文のレビューから，新たなベンチマークと称して，対応のない比較の場合は $d = 0.40$ 程度を「小」，0.70 程度を「中」，1.00 以上を「大」とみなす，といった提案をしているが，これは効果量の値を解釈するひとつの例とはなるものの，必ずしもこのようなベンチマークを利用することが推奨されるとは限らない。詳しくは，草薙（2015）の議論も参照のこと。
5. LET に掲載されている研究論文において，量的手法，その中でも統計的検定が，現在のように研究方法の主軸となったのは，主に 2000 年代に入ってからである。そのため，LET の第 38 号をはじめとして，その 12 年後にあたる第 49 号までを本研究の対象とした。

6. χ^2 検定は，原理的には χ^2 分布をもちいる検定のすべてを指すが，本稿が対象とした検定は，頻度分布に関する適合度検定である。また，事後検定は非常に少数であったため対象としなかった。
7. 五数要約は，最小値（minimum），第一四分位数（first quartile, Q_1），中央値（median），第三四分位数（third quartile, Q_3），最大値（maximum）をもってデータ要約すること，またはこれらの値の総称である。
8. η^2 の定義式は分母を全体の平方和として，η_p^2 は，誤差の平方和と要因の平方和の和とする。このことから η^2 と η_p^2 の関係は相関係数と偏相関係数の関係と同様であるととらえることができる。
9. ブートストラップ法は，再標本化法（resampling method）のひとつでもある。例えば得られた標本（元標本）から 1,000 回など莫大な回数，元標本からケースを再抽出（resampling）することによって母数の推定などをおこなう。また，ブートストラップ法は乱数をもちいた統計手法に性質が似通っているため，モンテカルロシミュレーション法のひとつともみなすことができる。外国語教育研究を背景としたブートストラップ法についての概説には草薙（2014a）がある。
10. ベイズ統計学は，ネイマン・ピアソン流の統計学（頻度主義ともいわれる）とは一線を画す確率論（ベイズ確率）にもとづく統計学である。頻度主義の下では，母数は未知であってもひとつであるととらえるが，ベイズ主義では母数自体が分布をなすととらえることができる。たとえば，通常の頻度主義では信頼区間を「a から b の範囲に母数が入る確率は 95%である」というように考えることができないが，ベイズ主義の下でこれは誤りではない。
11. 外国語教育研究に典型的な小標本の研究の場合，効果量の推定区間は得てして広い値を取りやすい。そのような推定のもとでは，確かな学術的解釈をあたえることができないと考え，効果量の信頼区間を報告することがためらわれる場合もあるかもしれない。また，査読などの過程においても，報告された信頼区間の広さが，研究の質を下げるものと判断されかねない。しかしながら，不確実性を含めて検討する手法が，不確実性を無視した手法よりも優れているのは自明である。さらに，現在の外国語教育研究では，メタ分析などによって研究の成果を体系的に統合することができるため，ひとつの研究において決定的な判断ができないとしても，その研究の貢献は容易に否定できるものではない。以上の点をご指摘いただいた査読者に感謝いいたします。
12. 以下のアドレスからダウンロードできる。https://goo.gl/AWafOc
13. データの可視化は，技術的に多少の訓練を要するものであり，敷居が高いと感じる研究者は少なくないと考えられる。しかしながら，近年は R に関係するものをはじめとして，データの可視化についての書籍の出版も増え，研究会なども各地で開催されており，可視化に関わる技術について学ぶ機会は増えている。

参考文献

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th edition). Washington, D.C.: Author.

Champely, S. (2012). pwr: Basic functions for power analysis. R package version 1.1.1. http://CRAN.R-project.org/package=pwr

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edition). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Erdfelder, E., Faul, F., & Buchner, A. (1996), GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, and Computers*, *28*, 1–11.

Frick, R. W. (1999). Defending the status quo. *Theory & Psychology*, *9*, 183–189.

南風原朝和 (1995).「教育心理学研究と統計的検定」『教育心理学年報』*34*, 122–131.

南風原朝和 (2002).『心理統計学の基礎―統合的理解のために』有斐閣アルマ.

南風原朝和 (2014).『続・心理統計学の基礎―統合的理解を広げ深める』有斐閣アルマ.

平井明代 (2012).『教育・心理学系研究のためのデータ分析入門』東京図書.

平野絹枝 (2011). 「本学会における研究（1991–2010）のレビューと展望」『中部地区英語教育学会研究紀要』*40*, 307–314.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

草薙邦広 (2014a).「外国語教育研究におけるブートストラップ法の応用可能性」『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報告論集』*5*, 1–15.

草薙邦広 (2014b).「外国語教育研究と直交表を用いた実験計画―実験計画の効率化を求めて―」『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報告論集』*4*, 24–33.

草薙邦広 (2014c).「外国語教育研究における量的データの可視化―分析・発表・論文執筆のために―」『外国語教育メディア学会中部支部外国語教育基礎研究部会 2013 年度報告論集』53–70.

草薙邦広 (2015).「教育実践のなかで集団に対する処遇の結果を適切に解釈するための定量的方法―効果量の利用とその限界点―」『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報告論集』*7*, 45–82.

Larson-Hall, J. (2012). Our statistical intuitions may be misleading us: Why we need robust statistics. *Language Teaching*, *45*, 460–474.

Larson-Hall, J., & Herrington, R. (2010). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics, 31*, 368–390.

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, *65*, *Supplementary 1*, 127–159.

Loewen, S., & Gass, S. (2009). Research timeline: The use of statistics in L2 acquisition research.

Language Teaching, 42, 181–196.

Loewen, S., Lavolette, E., Spino, L. A., Papi, M., Schmidtke, J., Sterling, S., & Wolff, D. (2014). Statistical literacy among applied linguists and second language acquisition researchers. *TESOL Quarterly*, *48*, 360–388.

前田啓朗 (2000).「構成概念の妥当性の検証 : 日本の英語教育学研究における傾向と展望」『外国語教育評価学会研究紀要』*3*, 119–126.

前田啓朗 (2004).「因果分析の妥当性の検証 : 日本の英語教育学研究における傾向と展望」『日本言語テスト学会研究紀要』*6*, 140–147.

前田啓朗 (2010).「生徒・学生理解のための教育データ分析」『第 36 回全国英語教育学会大阪研究大会発表予稿集』605–607.

Mizumoto, A., Urano, K., & Maeda, H. (2014). A systematic review of published articles in *ARELE* 1-24: Focusing on their themes, methods, and outcomes. *ARELE*, *25*, 33–48.

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.

水本篤・竹内理 (2011).「効果量と検定力分析入門―統計的検定を正しく使うために―」『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 2010 年度報告論集』47–73.

永田靖 (2003).『サンプルサイズの決め方（統計ライブラリー）』朝倉書店.

Norris, J. M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning*, *65*, *Supplementary 1*, 97–126.

Norris, J. M., & Ortega, L. (Eds.). (2006). *Synthesizing research on language learning and teaching*. Philadelphia: John Benjamins.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*, 241–286.

大久保街亜 (2009). 「日本における統計改革―基礎心理学研究を資料として―」『基礎心理学研究』*28*, 88–93.

大久保街亜・岡田謙介 (2012).『伝えるための心理統計―効果量・信頼区間・検定力―』勁草書房.

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*, 655–687.

Plonsky, L. (2014). Study quality in quantitative L2 research (1990-2010): A methodological synthesis and call for reform. *Modern Language Journal, 98*, 450–470.

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, *61*, 325–366.

Plonsky, L., Egbert, J., & Laflair, G. T. (2014). Bootstrapping in applied linguistics: Assessing its

potential using shared data. *Applied Linguistics*, Advanced Online Publication. http://applij.oxfordjournals.org/content/early/2014/02/14/applin.amu001.short

Plonsky, L,. & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.

Rosenthal, R., & Rosnow, R. L. (1984). *Essentials of behavioral research: Methods and data analysis*. New York: McGraw-Hill.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology*, *58*, 646–656.

Sdlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect of the power of studies? *Psychological Bulletin*, *105*, 309–316.

杉澤武俊 (1999).「教育心理学研究における統計的検定の検定力」『教育心理学研究』*47*, 150–159.

鈴川由美・豊田秀樹 (2011). 「『認知科学』における効果量と検定力，その必要性」『認知科学』*18*, 202–222.

鈴川由美・豊田秀樹 (2012). 「“心理学研究”における効果量・検定力・必要標本数の展望的事例分析」『心理学研究』*83*, 51–63.

竹内理・水本篤（編著）(2014).『外国語教育研究ハンドブック（改訂版）』松柏社.

豊田秀樹 (2009). 『検定力分析入門―R で学ぶ最新データ解析―』東京図書