# 英語の文法処理研究における統計的仮説検定：帰無仮説を主張する処遇について

名古屋大学大学院　草薙 邦広

## 1．研究背景

### 1．1 はじめに

統計的仮説検定自体の手法上の限界点や，統計的仮説検定の誤用・誤解についての指摘は，心理学・社会学の分野に限らず枚挙に暇が無い（例：大久保・岡田, 2012; 南風原, 1995）。しかしながら，依然として国内外を問わず，英語などを対象とする外国語教育研究では母平均値の比較を行う $t$ 検定や分散分析などを用いた研究が圧倒的多数である。

本稿は，統計的仮説検定の使用自体についての嫌疑を問うものではないが，これまでの英語の文法処理研究においてしばしば見られる，帰無仮説を研究仮説とする議論による結果の解釈について，その問題点を指摘し，今後の研究方上の展望を述べるものである。

### 1．2 帰無仮説を研究仮説とする統計的仮説検定

2000 年代以降，自己ペース読み課題（self-paced reading task）や視線計測を行える実験環境が第二言語習得の研究者の間にも広く普及されるようになり，母語話者と学習者が特定の言語的条件を持つ刺激文を読解する際に，どのような読解時間の変化を及ぼすか，といった情報が得やすくなっている。それに伴い，現在では母語話者および学習者における，特定の語や句の読解時間を比較する手法が一般的になってきているといえよう。

具体的には，母語話者と学習者の文法処理過程や，構文解析の選好性などの違いを明らかにするために，母語話者の読解において見られる読解時間のパターン（平均読解時間の差）が，学習者の場合には見られないという論理を用いる場合がある。統計的仮説検定をこの論理に当てはめるならば，ある言語的要因（e.g., 数の一致の正用／逸脱）が母語話者の読解時間に統計的に有意な差を及ぼすが，学習者の読解時間に対して有意差を及ぼさない，というような研究仮説を設定することになろう。すなわち，分散分析か $t$ 検定やその他の諸検定かを問わずとも，学習者の読解時間の統計量において，統計的な帰無仮説

（e.g., 二つの標本における母平均値が等しい）を棄却しないことにより，研究上の主張を支持することになる。通常の統計的仮説検定の場合は，帰無仮説を任意の水準で棄却することにより対立仮説を支持するが，この点が上記の場合と根本的に異なる点に留意されたい。これら帰無仮説自体を研究仮説とする一連の研究は，Online Sensitivity Paradigm とでも呼ぶべき理論的に重要な知見を構築して来たといえよう。しかし，その統計分析的，および研究計画法上の問題点が論じられることは多くなかった。

次節からは，統計的仮説検定における誤謬や，近年我が国の外国語教育研究においても注目が集まってきている効果量，検定力，そして必要標本サイズといった基礎的な概念に触れながら，帰無仮説を研究仮説とすることの問題点とは何かについて述べていく。

## １．３ 第一種と第二種の誤謬，標本サイズ

統計的仮説検定が誤謬を犯す場合（検定の結果が事実として真ではない）には，一般的に二種類のタイプがあるとされている。一つ目の誤謬は，第一種の誤謬（type I error）と呼ばれ，帰無仮説が事実として真である場合に，統計的仮説検定が帰無仮説を棄却する場合を表す。一方，第二種の誤謬（type II error）は，帰無仮説が事実として偽である場合に，帰無仮説を正しく棄却しない場合を示す。

上記の学習者の読解時間における例を取り上げると，事実として学習者の読解時間（条件ごとにおける母平均値）に差が無い場合に，統計的仮説検定が読解時間に有意な差を検出することが第一種の誤謬であり，第二種の誤謬は，読解時間に（事実として）差が存在する時に検定がそれを検出しない（見逃す）ことを示す。

通常，統計的分析を用いる殆どの学術分野では，第一種の誤謬の可能性に対してより保守的であるべきと考えられている為，第一種の誤謬の可能性（通常 $\alpha$ で表され，有意水準や危険率と呼ばれる）をより低く設定するようにしている。外国語教育研究や第二言語習得研究では，$\alpha=.05$，ないし $\alpha=.01$ と設定して検定を行う場合が多い。

しかしながら，当然のこととして，第一種の誤謬（$\alpha$）と第二種の誤謬の可能性（通常 $\beta$ で表せられる）は本質的に二律背反的な関係にあると考えられる（豊田，2009）。その為，帰無仮説を研究課題とする研究の場合，第二種の誤謬についてより保守的な態度を取るべきであるのに対して，そのような方策は容易には得がたい。ここで，一般に，標本サイズが大きくなれば大きくなるほ

ど，統計的仮説検定は有意差を検出しやすくなるという傾向があることにも留意すべきであろう。$t$ 値などの検定統計量もまたそうであるが，検定統計量を固定とした場合においても，自由度が高くなればなるほど，対応する $p$ 値は小さくなる。つまり標本サイズの少なさは，第二種の誤謬の可能性を高める。しかしながら，研究運用上の難しさもあろうが，当該の英語における文処理研究が殊更大きな標本サイズを持つことは多くない。

## １．４ 効果量，検定力，そして必要標本サイズ

国内の外国語教育研究においても，統計的仮説検定が標本サイズに依存しやすい性質を指摘し，効果量や検定力分析の普及につとめる論文や書籍も見られるようになった（e.g., 水本・竹内, 2008, 2011）。

効果量には様々な種類がある（Cohen, 1988; 水本・竹内, 2008 を参考にされたい）が，二群の平均値の差を評価する Cohen's $d$ がその代表のひとつであろう（この効果量にも様々な計算方法がある。詳しくは，豊田, 2009; 水本・竹内, 2011 を参考のこと）。一般的に用いられる，二標本のサイズが同じであり，標本間のデータの相関を考慮しない場合における定義式を以下の(1)に示す。参考に検定統計量 $t$（対応のあるデータに対する検定で用いる場合）の定義式を（2）に示す。これらの式から，$d$ は検定統計量 $t$ とは異なり標本サイズに依存せず，単純に共通の標準偏差（二群のばらつき）によって標準化された平均値の差（standardized mean difference）であると理解できよう。

$$(1) \qquad \text{Cohen's } d = \frac{(\text{標本１の平均値} - \text{標本２の平均値})}{\sqrt{\frac{\text{標本１の標準偏差}^2 + \text{標本２の標準偏差}^2}{2}}}$$

$$(2) \qquad t = \frac{(\text{標本１の平均値} - \text{標本２の平均値})}{\sqrt{\frac{\text{標本１の標準偏差}^2 + \text{標本２の標準偏差}^2}{\text{標本サイズ} - 1(\text{自由度})}}}$$

標本サイズに依存しない，このような種類の効果量に基づいて議論を行うことは，帰無仮説を研究仮説とする研究においても有効であろうが，これまでの研究で必ずしも効果量の報告が徹底されてはいない，という問題もある。また，効果量の算出（$d$ の場合）に必要な標準偏差などの記述統計が報告されていない例も散見される（しかし APA マニュアルには報告を義務付ける言明がある，American Psychological Association, 2009）。

効果量と，そして標本サイズ，有意水準が得られれば，その統計的仮説検定における第二種の誤謬を犯さない確率（検定力，1-β と記される）を求めることが出来る。これらの値は，上記の変数の内，三種類があれば，残りのひとつが求まる関係にある。このような性質を利用して，事前分析として，効果量が既知の場合に，任意の検定力と有意水準に対応する標本サイズを求めたり，実際の統計的仮説検定の結果から（所与の効果量，任意の標本サイズおよび有意水準），実際に行われた検定における検定力を求めることが可能である。これを事後分析と呼ぶ。これら一連の分析を検定力分析（power analysis）と呼ぶ。

固定的な水準として捉えるべきではなかろうが，一般に検定力は 0.80 以上が望ましいとされる（Cohen, 1988）。帰無仮説を研究仮説とする研究の場合では，第二種の誤謬を犯さない確率は，主張が事実として偽である時に，統計的仮説検定が偽と判断することに他ならないのであるから，検定力が低い検定は，たとえ帰無仮説が棄却されたとしても，主張に対する支持に根本的な弱さを抱えることになる。つまり，帰無仮説を研究仮説とする場合には，検定力を十分確保するような検定デザインが望ましい。例えば，第一種の誤謬と同程度である.95 に設定するべきという見解もあり得よう（例；杉澤，1999；大久保・岡田，2012）。

これまで国内外の心理学研究において，当該分野の研究における検定力を分析する事例研究がいくつかなされてきた（e.g., 杉澤, 1999; 鈴川・豊田, 2012）。しかしながら，それらの殆どは，通常望ましいとされる 0.80 の検定力に満たない研究が非常に多いことを指摘している。標本の大きさなどには現実的な制約が加わり，また逆に高すぎる検定力も問題ではあるものの，概して望ましい状況とはいえまい。更に，杉澤（1999）は『教育心理学研究』における帰無仮説を研究仮説とする研究の事例を取り上げ，その検定力が他の研究に比べて著しく低い事も指摘している。

このように，特に帰無仮説を研究仮説とする研究においては，検定力の確保が重要な課題であるのにも関わらず，本稿が対象とする一部の英語の文法処理研究において，検定力の観点がこれまで十分に踏まえられているとは言い難い。次章からは，帰無仮説を研究仮説とする英語文法処理研究の事例を見ながら，その効果量，検定力について考察を加えていく。

## ２．　検定力分析の事例：Jiang（2007）を対象に

Jiang（2007）は，第二言語習得研究において自己ペース読み課題を用いた代表的な研究のひとつである。当該の研究の実験参加者は 26 人の高熟達度英

語学習者（中国語を母語とする）および26人の英語母語話者，計52人であった。刺激となった文法項目は英語の動詞下位範疇化情報の逸脱（e.g., *The mayor promised to [offer/keep] the return advisor a better position soon*）および，数の一致の逸脱であった（e.g., *The child was watching some of the [rabbit/rabbits] in the room.*）。各種類16文ずつ（文法性によるカウンターバランスは取られている）についてPC上で移動窓式の自己ペース読み課題を実施し，正文・誤文条件間の読解時間を計測し，読解時間を単語単位で比較した（対応のある *t* 検定）。

結果，母語話者は動詞下位範疇化情報，数の一致，いずれの条件でも読解時間の差が有意であったのに対して，学習者の場合は，動詞下位範疇化情報のみで統計的な有意差が見られ，数の一致における読解時間の差は有意ではなかった。このことから，学習者は，数の一致に関して母語話者ほどの統合的知識（integrated knowledge）は持たず，言語項目によって知識の統合における過程は選択的である（selective integration）と主張している。

さて，ここで，主張の軸となる学習者・数の一致に関する読解時間の比較について詳細に見てみる。有意差が検出されなかった当該の対象領域の読解時間の報告を表1に参照する。

表1. Jiang（2007）の自己ペース読み課題における読解時間の一部（Jiang, 2007, pp. 17-18より参照；単位はミリ秒）

| 対象領域 | 3（誤りが明らかになる語） | | 4（次の語） | |
|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* |
| 正文条件 | 422 | 61 | 371 | 61 |
| 誤文条件 | 415 | 71 | 388 | 90 |

注：学習者（$n = 26$）毎に集計した平均値である。この二つの対象領域に有意差は無い。

このニペアにおける読解時間の差についての効果量が論文中では報告されていないが，所与の *M* と *SD* から著者がCohen's *d* を求めたところ（対応のあるデータであるため，便宜的に相関係数を0.50と仮定した），対象領域3：$d = 0.11$，対象領域4：$d = 0.18$ となり，Cohenの基準（Cohen, 1989; 水本・竹内, 2008）によるところの小程度の効果量といわれる0.20よりも小さいことがわかる。

効果量は前述の通り標本サイズに依存せず，標本サイズが検定の結果に及ぼす影響が無いものと捉えられるため，単純な効果量の解釈の上では「数の一致の誤りに対して鈍感である」というJiangの主張に対する根本的な批判とはな

らない。

しかし一方，上記の効果量に対する事後の標本検定力（α = .05，$N$ = 26 として計算した）を求めると，対象領域 3 において，Power = .08 であり，対象領域 4 では，Power = .14 となる。これは帰無仮説を研究仮説とする場合に望ましいと考えられる水準（例として power = .95）に遠く及ばないだけでなく，Cohen による検定力の一般的な基準である.80 にも満たない。

仮に，任意の効果量，有意水準，検定力を満たす標本サイズ（明日への分析と呼ばれる：豊田, 2009）を試算すると，$d$ = 0.11, α = .05, power = .80 とした場合，$N$ = 651 となり，$d$ = 0.18，および同値の α と検定力の場合では，$N$ = 245 となる（図 1 に α = .05，Jiang の結果から推定した効果量，対応のある両側 $t$ 検定と定めた場合における検定力と標本サイズの関係を示す）。このように今回の Jiang の実験によって得られた標本効果量と実験計画において，望ましいとされる 0.80 の検定力を到達するためには，通常の文処理研究では見られない大規模の実験協力者数が必要となる。しかしこのような規模の実験は実際性，および研究の費用便益分析的観点から望ましいとは言い難いだろう。

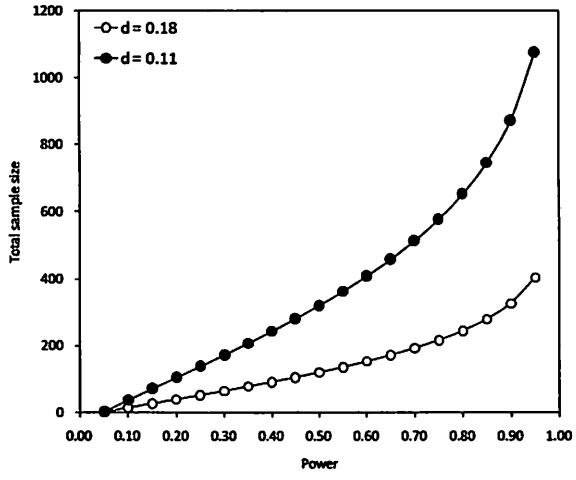


図 1. α = .05 および所与の効果量（$d$ = 0.18, $d$ = 0.11）における
検定力と標本サイズの関係

この事から，近年の英語文処理研究に頻繁に見られる「学習者が文法的な逸脱に対して鈍感性を示す」という知見が，従来の統計的仮説検定において読解時間に統計的な差がないというと分析のみに強く依拠することは，検定力分析の観点を踏まえると明らかに問題があると考えざるをえない。

## ３．検討すべき他の手段

ここからは，実際に上記の問題に対して講じ得る統計的分析方法の展望を述べる。本稿では，（１）確率分布に基づく信頼区間の推定，（２）ブートストラップ法を用いた種々の分析，（３）ベイズ因子に基づく決定法の三つに触れる。紙幅の都合上，各種手法の仔細や原理を述べることは本稿の及ぶところではないが，外国語教育研究における手法上の検討に対する一助になればと，上記のデータを例として上げながら概観する。

### ３．１ 信頼区間の推定による議論

標本データを特定の確率分布に見立てる場合，分布の特徴から，真値に対応する値の区間を推定することができる。一般に母数における任意の統計量の推定区間を信頼区間（confidence interval: CI）と呼び，標本値におけるものを予測区間（prediction interval: PI）と呼ぶ。通常 95%や 99%を基準として設ける。母平均値の信頼区間を例とするならば，正規分布に従うと考えられるデータであれば，標本平均値は母平均値の点推定値と同値であるわけだが，標本の分散と標本サイズから，標準誤差が求まるため，母平均値を中心に任意の区間の推定を行うことができる。[1]

（母平均値の）信頼区間による検討の長所は，平均差の検定や効果量による議論と異なり，実測値におけるスケール上で意思決定を行うことができる点である。Jiang による当該データを用いると母平均値の 95%信頼区間は以下のように求まる（表 2）。また，信頼区間は通常エラーバーを用いて図 2 のように示すことが多い。

表 2. Jiang（2007）のデータにおける母平均値の信頼区間の計算例 1

| | 対象領域 3 | 対象領域 4 |
|---|---|---|
| 正文条件 | [397, 447] | [346, 396] |
| 誤文条件 | [386, 444] | [352, 424] |

注．水準を 95%，計算は確率分布（正規分布）に基づく計算方法を用いた。

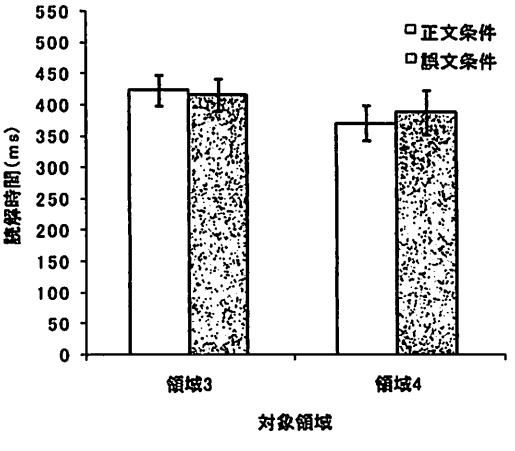


図 2. 各条件における平均値と母平均値の 95%信頼区間

このように，信頼区間を検討するとどちらの条件間においても，信頼区間が大きく重複していることが分かり，実際のスケール上においても大きな読解時間の差がないだろうという統計的推測を容易に行うことができる。また，信頼区間は測定の精度も表すため，任意の誤差（ないし信頼区間）を目標として標本サイズを決定するという方法も存在する。このような分析手法は総じて統計的仮説検定の二値的判断よりも情報力が多いことはいうまでもないだろう。

### 3．2 ブートストラップ法を用いた方法

再標本化統計の一種であるブートストラップ法は，標本値の復元抽出を用いた計算シミュレーション（ノンパラメトリック・ブートストラップ法），または標本から得られたパラミターの推定値によって与えられる確率分布に従う乱数を用いたシミュレーション（パラメトリック）を，繰り返し莫大な回数行うことによって，確率分布に基づいた従来の統計的手法の代用を試みるものである（汪他，2003）。外国語教育研究においても，近年頑健な分析手法として紹介されるようになってきている（e.g., Larson-Hall & Herrington, 2010）。

ブートストラップ法は種々の統計量に対して適用することができるが，上記の信頼区間についても，確率分布を使用せず，標本の実測値に基づくシミュレーションによって計算することができる。仮に，パラメトリック・ブートストラップ（$B$ = 1,000，ブートストラップ標本サイズを，元標本と同値である $n$ = 26，パーセンタイル法と定めた場合）[2] を用いて，上記のデータの 95%信頼区間を求めると表 3 のようになる。この計算はシミュレーションを経るアルゴリズムのため，施行によって数値は微小に変化することに留意されたい。

表 2. Jiang（2007）のデータにおける母平均値の信頼区間の計算例 2

| | 対象領域 3 | 対象領域 4 |
|---|---|---|
| 正文条件 | [397, 444] | [348, 394] |
| 誤文条件 | [390, 444] | [354, 422] |

注．水準を 95%としてパラメトリック・ブートストラップ法（$B$ = 1,000, bootstrap $n$ = 26，正規分布を指定）を用いた。しかし本来，反応時間データに対して正規分布を適用するのは望ましいとはいえない。[3]

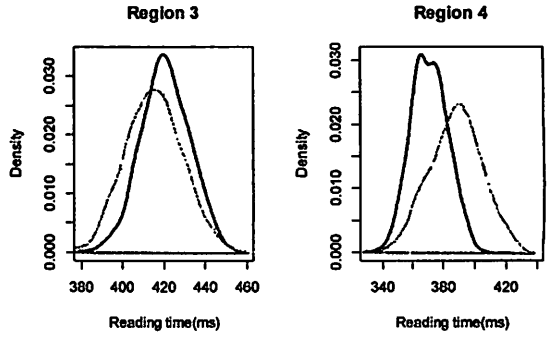


図 3．ブートストラップ法による母平均値の確率分布，黒線が正文条件の確率密度曲線，灰が誤文条件である。

また，各ブートストラップ標本の平均値について度数分布（および確率密度曲線）を求め，それを図示することによって，母平均値の推定をより視覚的に行うことができる（図 3)。可視化した母平均推定値の確率分布を見ると，対象領域 4 では，やや誤文条件の分布が正の方向に歪んでいる傾向が分かる。

ブートストラップ法は汎用性が非常に高いため，原理的には，平均差，効果量，*t* 値，有意確率などといった指標についてもブートストラップ信頼区間を求めることができる。他にも，これを応用することによって，効果量が同じ符号を取る確率（ベイズ的予測を用いる *Prep* も参照のこと，例として大久保・岡田，2012；村山，2010；Kelleen, 2005）の計算機的シミュレーションなども容易に行うことが可能であるため，統計的仮説検定の代用として期待するところが大きいだろう。

しかしながらブートストラップ法は元標本に大きく依存する傾向があり，その他にも種々の手法的問題点が指摘されている。それらの点については注意が必要であろう。

### 3．3 ベイズ因子を用いた方法

近年，心理学などの分野を中心に，ベイズ統計学的アプローチにより，統計的仮説検定を代用しようとする試みがある（e.g., Dienes, 2008; Rouder et al, 2009; Morey & Rouder, 2011)。

ベイズ因子の最大の特徴は，二つの任意の仮説の確からしさの比率を検証することである。つまり，従来の帰無仮説を棄却することのみによって対立仮説を採択する論理とは根本的に異なる。また，ベイズ式の信頼区間であるベイズ信用区間という区間推定法もある。

本稿では手法の名前を挙げる程度に留まるが，今後このような手法が外国語教育研究でも発展するだろうと考えられる。また近年，Web ベースの計算サービス，または統計解析環境である R 対応のパッケージが複数公開されているため，今後益々手法の発展が期待されるところである。[4]

## 4．結論

本稿は，近年英語の文法処理研究で特に多く見られる，帰無仮説を棄却しないことを主張の根拠とする一連の研究の手法的問題点を，主に統計的仮説検定自体が持つ構造的弱点，および検定力の観点から考察した。また，手法上の代替案として，信頼区間を用いた検討，ブートストラップ法，そしてベイズ因子による統計的意思決定法を紹介し，その展望を述べた。

本稿が，今後，同種の研究課題および研究パラダイムにおける分析手法選択の一助となることを願う。

## 注

1. 母平均値における95%信頼区間は，平均値±1.96×標準誤差とされる場合が多いが，これは自由度が∞である場合の計算である。自由度を考慮する場合，自由度と確率に対応する *t* 値の逆数と標準誤差の積を用いるとよい。
2. *B* はブートストラップの試行回数（ブートストラップ標本数）を表す。ブートストラップ標本サイズは推定区間に大きく影響を及ぼすが，通常元標本と同値を用いる場合が多い。推定方法には，パーセンタイル法の他にもベーシック法，BCa法，パーセンタイル *t* 法など様々なものがある。
3. 反応時間データは通常，ワイブル分布，Ex-Gaussian 分布ないし混合正規分布を用いた分析がなされる場合が多い。
4. 代表的な web ベースのツールとして，*Bayes Factor Calculators*（http://pcl.missouri.edu/bayesfactor）また，統計解析環境 R における解析用パッケージとして，*BayesFactor*（http://bayesfactorpcl.r-forge.r-project.org/）などがある（2013 年 10 月アクセス）。

## 参考文献

American Psychological Association. (2009). *Publication manual for the American Psychological Association* (6th ed.). Washington, DC: Author.

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

Dienes, Z. (2008). *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference*. Palgrave Macmillan.

Jiang, N. (2007). Selective integration of linguistic knowledge in adult second language learning. *Language Learning*, *57*, 1–33.

Killeen, P. R. (2005). An alternative to null hypothesis significance tests. *Psychological Science*, *16*, 345-353.

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.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406–419.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225–237.

汪金芳他（2003）『計算統計I—確率計算の新しい手法（統計科学のフロンティア11）』東京：岩波書店

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』東京：勁草書房

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（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.

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

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.

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

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—」『より良い外国語教育研究のための方法：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. Retrieved from http://www.mizumot.com/method/mizumoto-takeuchi.pdf

村山航（2010）「Prep について」Retrieved from http://www4.ocn.ne.jp/~murakou/prep.pdf