複数の項目やテストにおける検定の多重性 : モンテカルロ・シミュレーションによる検証

doi: 10.24539/let.46.0_1

Language Education & Technology, 46 (2010)
pp. 1–19

複数の項目やテストにおける検定の多重性
—モンテカルロ・シミュレーションによる検証—

Multiple testing of several items or tests:
A Monte Carlo simulation study

水本　篤
流通科学大学

MIZUMOTO, Atsushi
University of Marketing and Distribution Sciences

This paper aims to highlight the problem of multiple significance testing with several dependent variables (i.e., items or tests). In many research papers, researchers report the results of multiple significance testing without realizing they are committing Type I error, in which it can be erroneously concluded that there is a statistically significant difference, when in fact there is no statistical difference. In order to address this problem, a series of Monte Carlo simulation studies were carried out. Five artificial sets of dependent variables for two groups of subjects were generated in the simulation. Three types of data sets which varied in their degrees of intercorrelations (r = .00, r = .50, r = .95, respectively) were then compared. The results indicate that multiple significance testing, with several dependent variables, inflate Type I error, and thus caution should be exercised to control the experimentwise error rate. Implications for the strategies for controlling Type I error rate are then discussed.

Keywords: 統計的検定，検定の多重性の問題，第 1 種の誤り，多変量分散分析

1. はじめに

1.1　研究の背景

研究論文において使用される統計的検定では，3 つ以上のグループの平均値を比較する場合には，t 検定を繰り返すのではなく，分散分析（ANOVA）を行い，全体として有意差があった場合には，多重比較でペアごとの検定が行われる。これは，t 検定を繰り返して行った場合には，本当は差がないのに統計的に有意

差があると判定してしまう「第 1 種の誤り」の確率が高くなってしまうからであり，このような説明は統計データ分析の入門書などでは必ず加えられている (例えば，磯田, 2004 など)。そのため，このようなケースにおける多重比較は研究論文でも適切に報告されていることが多い。表 1 は，一般的に正しく理解されているそのような方法を示した例である。この表では，一元配置分散分析を行ったあとで, どのグループの組み合わせに統計的に有意な差が見られるかを調べるために，グループ A とグループ B，グループ A とグループ C，グループ B とグループ C の 3 回の検定（多重比較）を行っていることを示している。

表 1　一般的によく知られている分散分析後の多重比較（3 回の検定）の例

	グループ A[a]	グループ B[a]	グループ C[a]	*F*	*p*
リーディングテスト	65.34 (17.63)	58.83 (12.01)	50.82 (12.82)	10.24	.000

Note. [a]*n* = 40;Mean (*SD*)

しかし，いくつかの質問紙項目（もしくは下位尺度）やテストのような複数の従属変数を，複数のグループ間で比較するような場合でも，検定を繰り返すことによって第 1 種の誤りを犯す確率が高くなるという事実はあまり知られていないため，研究論文でそのような結果を提示する際に，有意水準の調整をせずに報告されている例が散見される。そのような理由から，前田 (2008a) が紹介しているように，論文投稿者がジャーナルの査読者に第 1 種の誤りの可能性を指摘されながらも，何が問題なのか理解できないこともある。表 2 がそのような例であり，リーディングテストで分散分析を 1 回行い，リスニングテスト，スピーキングテストでも，それぞれで分散分析を 1 回ずつ行っているため，計 3 回の検定を同じサンプル（実験参加者）で行っていることになる。ちなみに，この例では分散分析（ANOVA）の繰り返しになっているが，他の検定（例えば *t* 検定，χ^2 検定など）でも同様にこのケースに当てはまる。[1]

このように，検定を同じサンプルに対して何度も繰り返すことによって，第 1 種の誤りの確率を高くしてしまうことを，「検定の多重性の問題」と呼び (足立, 1998)，本稿では特に表 2 の例のような，従属変数がいくつかある場合に生じる検定の多重性の問題を取り上げる。

表 2　従属変数が複数ある場合に起こる検定の多重性の問題（3 回の検定）の例

	グループ A[a]	グループ B[a]	グループ C[a]	*F*	*p*	
リーディングテスト	65.34 (17.63)	58.83 (12.01)	50.82 (12.82)	10.24	.000	⇦
リスニングテスト	59.13 (9.76)	57.21 (7.61)	53.77 (7.46)	4.39	.017	⇦
スピーキングテスト	64.33 (14.54)	57.17 (13.61)	53.61 (19.81)	4.53	.013	⇦

Note. [a]$n = 40$; Mean (*SD*); 矢印は検定の回数を示す。

1.2　統計的検定と検定における 2 種類の誤り

いくつかの従属変数（質問紙の項目やテスト）における検定の多重性の問題を理解するために，統計的検定とそれに付随する 2 種類の誤りについて理解しておかなければならない。まず，検定は「統計的に有意な差がある」ということを示すために行うため，はじめに「差がない」という帰無仮説（null hypothesis）を設定する。そして，慣例として，検定結果が $p < .05$（5 %以下）であった場合には，100 回中で 5 回以下の低い確率で偶然起こることを表すので，「差がないということはない」と考えられる。そのため，設定しておいた帰無仮説を棄却し「差がある」という対立仮説（alternative hypothesis）を採択し，「統計的に有意な差がある」という判断を下す。つまり，有意確率と呼ばれる p 値は確率を表しているので，逆に p 値が .05 を超えた場合には，差がないという仮説に対する確率が高くなるため，「統計的に有意な差があるとはいえない」という判断になる。[2]

ここで注意しなければならないのは，$p < .05$ で有意な差があると判断しても，その判断が誤りである確率が p 値と同じだけあるということである。例えば，$p = .05$ であったとしても，母集団の性質の推定を行ったときに，20 回に 1 回は判断が外れる可能性があることを認めているのが統計的有意差検定の考え方である。このような理由から，α で表される有意水準は「実際には有意差がないのに有意差がある」と判断してしまう第 1 種の誤り（Type I error）を犯す確率を表しており，「危険率」とも呼ばれる。そして，その反対のパターンで，「実際には有意差があるのに有意差がない」と判断してしまうのが，第 2 種の誤り（Type II error）を犯す確率（β）である。第 1 種の誤りを犯す確率（α）は，統計的検定

の慣例として，α = .05 がよく用いられ，第 2 種の誤りを犯す確率（β）は，β = .20 を保つのが望ましいと提案されていている (Cohen, 1988)。[3]

検定におけるこれら 2 つの誤りに加えて，「検定力（power)」（または検出力）も大切な概念になる。検定力とは,「有意差を見つける力」(磯田, 2004, p. 48), あるいは「母集団において差があるとき，サンプルにおいて有意な結果が得られる確率」(南風原, 2002, p. 143) である。サンプルサイズを大きくすれば検定力は高まるが，逆に，検定力が強すぎる場合には，実質的な差がなくても，統計的に有意な差を検出する可能性がある。検定力は 1－β で定義され，β = .20 の場合，1－0.2 で 0.8 になる。検定力が 0.8 ということは，実際に有意差があるときには, 80 % の確率でそれを検出できることを意味している。また, Cohen (1992) は,「0.80 以下の検定力の場合には, 第 2 種の誤りを犯す可能性が高くなる」(p. 156) としている。表 3 は，これらの関係をまとめたものである。

表 3　統計的有意差検定における 2 種類の誤りと検定力

	差がないと判断 （帰無仮説を採択）	差があると判断 （帰無仮説を棄却）
本当は差がない	正しい判断（1－α）	第 1 種の誤り（α）
本当は差がある	第 2 種の誤り（β）	正しい判断（1－β）[検定力]

Note. 小野寺・菱村 (2005, p. 60) を引用，一部改変。

検定は同じサンプルに対して 1 回だけ行うことを前提に，α を 0.05 に設定しているので，前述の表 1 のように他群を比較するときだけでなく，表 2 のようにいくつかの従属変数（質問紙の項目やテスト）がある場合にも，検定を繰り返している場合には，α = .05 でコントロールしていたはずの，第 1 種の誤りを犯す可能性が高くなってしまう。n 回の検定の繰り返しによって第 1 種の誤りを犯している確率（α）は，理論的には次の式で求めることが可能である (足立, 1998; 山田・杉澤・村井, 2008)。

$$1 - (1-\alpha)^n$$

例えば，検定を 3 回繰り返した場合には，α = .05 であるとすると 0.143 [= 1 －

$(1- 0.05)^3$] になるので，第 1 種の誤りを犯している確率は 14.3 % であると考えられる (靜・竹内・吉澤, 2002)。つまり，$\alpha = .05$ の有意水準を設定して判断をするはずが，その値を超えてしまい，有意水準をコントロールできていないという結果となる。上述の計算式から，検定の繰り返しが増えれば増えるほど，この確率は高くなることは明らかであり，足立 (1998) で例に挙げられているように，質問紙を用いた論文で 100 項目に対して検定が繰り返し実施された場合には，第 1 種の誤りを犯す可能性が 99.4 % にもなってしまうため，ほとんどどこかの項目において誤って（たまたま）有意であるという結果になってしまう。

このような理由から，表 1 の例のような分散分析後に多重比較を行う場合には，テューキー（Tukey）の方法，シェッフェ（Scheffé）の方法，ボンフェローニ（Bonferroni）の方法，ダネット（Dunnet）の方法（詳しくは，永田・吉田, 1997 などを参照）のような手法を用いるのが一般的であり，論文において報告されている結果に問題がある場合は少ない。しかし，表 2 のように，いくつかの従属変数があり，一度に検定を行った場合には，SPSS（2009 年 4 月より PASW に名称変更）をはじめとする統計ソフトでは，この問題について明示されていない結果が出力されるため，研究者が第 1 種の誤りに気づいていない場合が多い。

前田 (2004) でも指摘されているように，このような状況は，コンピューターやソフトウェアの発展により，より身近になった（複雑な）分析手法を誤用に気づかずに使用していることと，それに伴い，先行研究の誤った手法の模倣を繰り返しているために起こっていると考えられる。そのような分析手法の誤用について，石井 (2005) は次のように述べている。

> 先行研究でこんな分析方法を使っていたから自分もそれと同じ分析方法を使うというのも時として危険です。論文発表された研究の中には，おかしな分析を行っているものもたくさんあるからです。先行研究のまねをした，おかしな分析を行っている研究が再生産されているというのは，悲しいことですが現実に起こっていることです。この繰り返しを断たないとその領域の学問は発展しませんし，科学と呼ぶに値しません。(p. 6)

このような考え方からも，外国語教育学が学問分野としてさらに確立していくためには，先行研究で間違っている可能性がある分析・結果の提示方法について，分野に携わる研究者それぞれが正しい方法を学んでいくことが大切である。

ゆえに，本稿ではいくつかの質問紙項目やテスト（複数の従属変数）がある場合に，検定を繰り返すことによって第 1 種の誤りを犯す確率が高くなるという多重性の問題に注意を向けるために，シミュレーション研究を行って問題点を

確認していく。

1.3 研究の目的

本研究では，前節までで説明したような，いくつかの従属変数がある場合に検定を繰り返すことによって第 1 種の誤りを犯している確率がどのように変化するかということを，モンテカルロ・シミュレーションと呼ばれる手法により検証する。そして，その結果から，第 1 種の誤りを犯す確率（α）をコントロールするにはどうするべきかという対処方法について考察を加えることを目的とした。

2. 方法

データ解析環境 R を用い[4]，従属変数（テストや質問紙の項目）が 5 つある場合に *t* 検定を繰り返すことにより第 1 種の誤りを犯す確率を，モンテカルロ・シミュレーションを使って計算した。モンテカルロ・シミュレーションとは，「乱数を用いたシミュレーションを何度も行うことにより，考えている問題の近似解を得る計算方法」(舟尾, 2005, p. 80) のことである。[5]

外国語教育学では Meara (2005) に見られるようなシミュレーション研究を行うことは多くないが，心理学などの研究分野ではよく用いられており (栗田, 1999 など)，シミュレーションを行うことで，実際のサンプルでは統制して分析しにくい点を厳密に検証できる点において有用である (Excel を利用したシミュレーションは 靜, 2007 に詳しい)。今回のシミュレーション実験に使用したデータは，舟尾 (2005)，山田・杉澤・村井 (2008)[6] などを参考に，以下の手順で発生させたものである（具体的な R のコマンドは Appendix A を参照）。

(1) *t* 検定の前提条件を満たすように，正規分布している母集団からの無作為抽出を行い，分散の等しい 2 群（平均 50，標準偏差 10）のデータを 5 つの従属変数分発生させる。2 群のサンプルサイズは G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007) を使い，両側検定，中程度の効果量（*d* = .50）[7]，α = .05（後述のボンフェローニの方法 [.05/5] により α’ = .01 に補正）であった場合には，検定力が Cohen (1988) によって推奨されている 0.80 以上（0.82）となる，各群 100 名とした。

(2) 2 群は母平均が同じ（差がない）設定なので，検定を行ったとしても正しく行われている場合は「有意差はない」という判定になるはずであり

（帰無仮説が正しい),「有意差はある」という結果になった場合には第 1 種の誤りを犯していることになる。そこで，それぞれの従属変数で *t* 検定を行い，有意になった数を基に，第 1 種の誤りを犯している割合を調べる。

(3) 上記 2 つの手順を 100 回, 1,000 回, 10,000 回, 100,000 回, 1,000,000 回と繰り返し，第 1 種の誤りを犯している割合（平均）を求める。

また，現実のデータでは，テスト間，もしくは項目間などの変数間に相関がある場合が多いと考えられるため (Field, 2005)，(a) $r = .00$（無相関），(b) $r = .50$（中程度の相関），(c) $r = .95$（とても強い相関）の 3 パターンの相関を持つデータを発生させることにより検討した（発生させた変数の相関係数の 1 例は Appendix B を参照）。

表 4 はシミュレーション実験で発生させたデータの 1 例である。シミュレーションでは, 正規分布している母集団から乱数を使って無作為にデータを抽出しているので，毎回同じ値になるわけではないが，すべてのデータが近似的に平均 50，標準偏差 10 になっていることがわかる。

表 4　シミュレーション実験で発生させたデータの例

	$r = .00$		$r = .50$		$r = .95$	
	Group A[a]	Group B[a]	Group A[a]	Group B[a]	Group A[a]	Group B[a]
Test 1	50.27	50.46	50.24	50.03	49.62	49.99
	(9.68)	(9.46)	(10.19)	(9.50)	(10.22)	(9.80)
Test 2	49.78	50.22	51.44	51.03	49.25	49.66
	(9.89)	(11.59)	(10.16)	(10.36)	(10.15)	(10.20)
Test 3	50.34	49.33	50.34	50.27	49.72	50.11
	(10.31)	(10.32)	(9.95)	(8.13)	(10.01)	(9.97)
Test 4	50.76	52.66	50.13	49.86	49.51	49.68
	(10.30)	(10.32)	(9.60)	(9.65)	(10.37)	(10.23)
Test 5	50.45	50.32	50.18	49.45	49.64	50.30
	(9.64)	(9.49)	(10.19)	(9.53)	(10.46)	(10.12)

Note. [a]$n = 100$; Mean (*SD*); Test 1～5 は従属変数を表している。表中の数値は乱数によって発生させているデータなので，毎回，$M = 50$, $SD = 10$ に近い値が得られるが必ずしも同じ値にはならない。

3. 結果

モンテカルロ・シミュレーションの結果を表 5 に示す。図 1 は同じ結果を図示したものである。これらの結果からわかるように，シミュレーションの回数が増えていくたびに数値が安定していき，10,000 回以上の繰り返しになると，ほぼ一定の値に収束している。また，第 1 種の誤りを犯している確率（α）は，従属変数間にとても強い相関関係（r =.95）がある場合には低くなり，相関がない（r = .00）場合には高くなることがシミュレーションの結果からわかった。しかし，従属変数間にとても強い相関関係（r =.95）がある場合でも，基準とされる .05 は超えていて明らかに問題であることがわかる。

理論的には，第 1 種の誤りを犯す確率は，Hair, Black, Babin, Anderson, and Tatham (2006, p. 400) でも説明されているとおり，5 つの従属変数があれば，変数間の相関が完全相関（r = 1.00）の場合での 5 パーセントから，無相関（r = .00）の場合の 23 パーセントの間のどこかの値になるとされている（無相関の場合には，前述の $1 - (1-\alpha)^n$ の式と同じ）。つまり，5 つの従属変数に対して t 検定を繰り返している今回のシミュレーションでは，理論値とかなり近い値が得られていることがわかる。そして，この値は（表 1 のような）グループ間での t 検定の繰り返しの際にも同じものであり，そのような場合と同様に，複数の従属変数の検定でも多重性の問題への注意が必要であることが確認された。そして，いくら従属変数間の相関係数を考慮に入れたとしても，有意水準として設定している 5 パーセントよりも大きな値になってしまうため，第 1 種の誤りを犯す確率のコントロールは行う必要があることが明らかになった。

表 5　モンテカルロ・シミュレーションの結果

繰り返し回数	r = .00	r = .50	r = .95
100 回	.220	.180	.080
1,000 回	.239	.188	.100
10,000 回	.219	.179	.089
100,000 回	.224	.184	.083
1,000,000 回	.226	.183	.084

Note. 数値はパーセントを表すため 1（100 %）が最大。

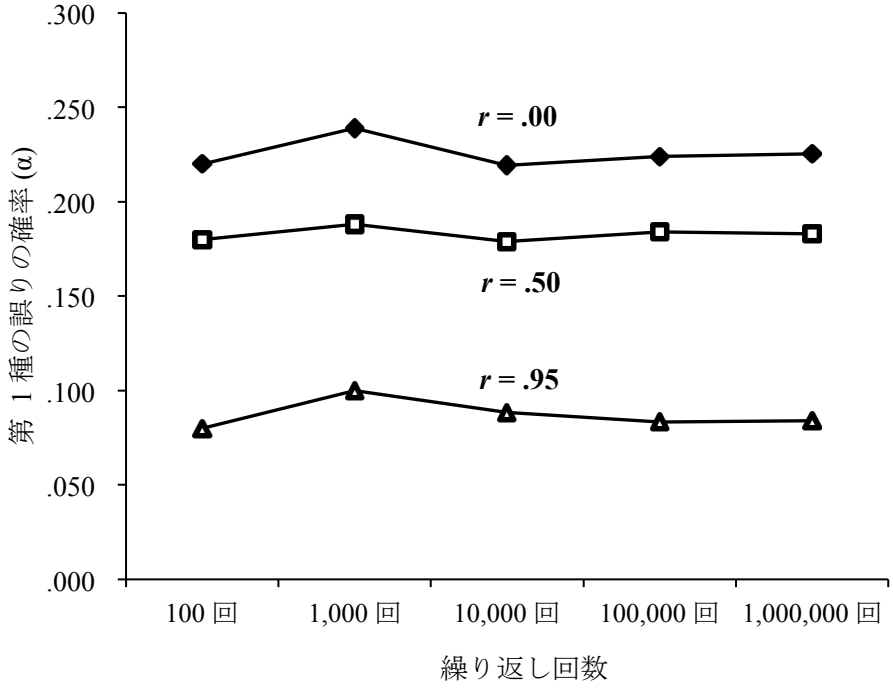

図 1. シミュレーション結果を図示したもの

4. 考察

モンテカルロ・シミュレーションの結果から，いくつかの質問紙項目やテスト（複数の従属変数）に対して検定を繰り返すことによって，第 1 種の誤りを犯す確率が高くなることが証明された。ゆえに，ここでは第 1 種の誤りを犯す確率（α）をコントロールする方法や，代替策についての考察を加えていく。

4.1 多変量分散分析（Multivariate Analysis of Variance: MANOVA）

本研究で取り上げた例のように従属変数がいくつかある場合には，*t* 検定や分散分析をそれぞれの従属変数について行うのではなく，多変量分散分析（MANOVA）を行うべきであるとする立場もある (多変量分散分析の詳細については，Field, 2005; 小野寺・菱村, 2005; Tabachnick & Fidell, 2006; Weinfurt, 1995 などを参照)。多変量分散分析では，従属変数間の相関を考慮し，全体として有意差があるのかの検定を行う (小野寺・菱村, 2005)。従属変数ごとに分散分析

（univariate ANOVA）を行った場合には，従属変数間の相関関係の情報が失われてしまうために，変数間に何らかの相関関係が想定される場合には，それぞれでの従属変数で分散分析（ANOVA）を行うよりも多変量分散分析（MANOVA）のほうがふさわしいとされている。

また，多変量分散分析を行ったあとに，それぞれでの従属変数ごとに分散分析を行うという手順が「第 1 種の誤りを犯す確率をコントロールしている」と考えられているため，第 1 種の誤りを防ぐ方法として現在では最も多く用いられている (Field, 2005; Hair et al., 2006; Weinfurt, 1995)。しかし，前田 (2008a) でも紹介されているとおり，この手順では，それぞれでの従属変数ごとで分散分析を行ったときに第 1 種の誤りを犯す確率をコントロールできていないと考えられる (詳細は Bray & Maxwell, 1982; Field, 2005; Harris, 1975; Huberty & Morris, 1989; Weinfurt, 1995 を参照)。具体的な例では，SPSS で多変量分散分析（MANOVA）を行ったときに出力される従属変数の分散分析（ANOVA）の結果と，それぞれの従属変数を使って別々に分散分析を行った結果は一致するため，Field (2005) は “The fact illustrates that MANOVA offers only hypothetical protection of inflated Type I error rates: there is no real-life adjustment made to the values obtained” (p. 602) と述べている。また，Weinfurt (1995) も，多変量分散分析を第 1 種の誤りの確率をコントロールする目的で使うことの是非について，次のようにまとめている。

> the MANOVA—multiple ANOVA technique does not protect the experimentwise alpha when there is a significant multivariate effect present. If protection against Type I error is the concern, there are more appropriate techniques available, such as the various versions of the Bonferroni correction. (p. 264)

多変量分散分析（MANOVA）は，前述のように複数の変数間の相関を考慮に入れて検定を行うことができるため，その点では有効な手法ではあるが，目的が第 1 種の誤りの確率をコントロールすることにあるのであれば，従属変数ごとに行う分散分析（ANOVA）でも，次に紹介するボンフェローニの方法などを利用すべきであろう。

4.2 ボンフェローニ（Bonferroni）の方法

ボンフェローニの方法は，有意水準（慣例では $\alpha = .05$）を，検定を繰り返した回数で割るというシンプルな方法ではあるが，本稿で取り上げている，従属

変数がいくつかある場合の第 1 種の誤りの確率を調整するために有効な方法である (Field, 2005; Weinfurt, 1995)。考え方としては，ある実験中において検定の繰り返しによって生じる第 1 種の誤りを犯す確率（experimentwise error rate）[8] が，全体で α = .05 を超えないようにコントロールするというものである。例えば，5 つの従属変数がある場合には検定を 5 回繰り返すことになるので，0.05/5 = 0.01 を α =.05 と同じものとして扱う。[9] つまり，*p* 値がそれぞれの従属変数における検定で .01 以下（$p < .01$）になったときには，「有意差がある」と判断する (磯田, 2004)。

このようにボンフェローニの方法は直感的にもわかりやすく，第 1 種の誤りの確率を調整するに有効であるが，検定の繰り返し回数が多くなれば，有意差の判定が厳しくなりすぎる（保守的である）ため，第 2 種の誤り（実際に差があるときに「ない」と判断する確率である β）が高くなってしまう。よって，ボンフェローニの方法は検定を行う回数が少ない場合に有効であると言われている (竹原, 2007)。永田・吉田 (1997) では，ボンフェローニの方法よりも検定力が高いホルム（Holm）の方法なども紹介されているが，検定の繰り返しが多すぎる場合には根本的な解決法にはならない。つまり，「たくさんの項目を含んだ質問紙でデータを収集したので，項目のすべてで検定してみる」というような考えでは，（探索的因子分析などで項目をまとめたとしても）検定の繰り返しが無意味に多くなってしまうので，統計的有意差検定を利用する研究では，実験計画の段階で第 1 種（および第 2 種）の誤りを犯す確率の調整に対する問題意識を持っておくことが必要であるだろう。

4.3 効果量（effect size）

有意差検定では，サンプルサイズ（サンプル数，あるいは被験者数）が大きくなればなるほど，統計的に有意であるという結果になりやすいという大きな問題がある。このため，同じ実験である検定を行ったところ，20 人では有意ではなく，200 人のデータの場合には有意になるということも十分にあり得る。このように，*p* 値はサンプルサイズに敏感なので，実質的な効果が大きいか小さいかについての情報を判断する際にはあまり有意義な指標ではない。また，「有意差とは確率的に差があると言っているだけで，差の意味までは説明してくれません。」(磯田, 2004, p. 34) という説明からも，有意差があったからといって，それがどれぐらいの差なのかということまではわからないのである。そのような理由から，サンプルサイズによって変化することのない指標である効果量（effect size）の報告が，論文においても必要であると定められている (American Psychological Association, 2001)。

芝・南風原 (1990) によると，効果量は，「測定単位にたよらない指標となっている。そのため，効果量を用いれば，単位の異なる変数を用いた研究の間でも，実験条件の効果の大きさを互いに比較することができる」(p. 118) と定義されている。つまり，有意差検定に依拠しない結果の解釈も可能になるため，単純に差の比較が目的であるのなら，第 1 種の誤りや第 2 種の誤りの問題を避けて，効果量による結果の提示・解釈も可能である (例えば，Koizumi & Katagiri, 2007)。また，効果量は，サンプルサイズを決定するために利用できる検定力分析（power analysis）で用いたり (Field & Hole, 2003; 村井, 2006)，いくつか複数の量的な研究結果を統合して，全体としての効果を検討する統計的分析手法であるメタ分析（meta analysis）にも用いられるため (例えば，In'nami & Koizumi, 2009)，論文においては必ず報告すべきである (効果量についての説明は Cohen, 1988, 1992; Kline, 2004; 水本・竹内, 2008 などを参照)。

4.4 平均値と標準偏差の吟味

効果量による結果の解釈に準じた考え方として，検定をまったく用いずに，平均値と標準偏差の吟味（またはグラフ化）によって，結果の解釈を行うことも可能である (磯田, 2008 や 前田, 2008b がその良い例である)。

そもそも，外国語教育学で主に扱われる学習者から得られたようなデータは，推測統計の前提である母集団からの単純無作為抽出を行ったとは見なせないものがほとんどであるため，手元のデータから母集団の性質を推定する際には，過度の一般化は控えるべきである (南風原, 2002)。また，介入の効果を測定するために，指導前と指導後におけるデータを用いる研究（プリ・ポストデザイン）も多いが，「平均への回帰」（または回帰効果 : regression to the mean）は普遍的な現象であるため，誤った結論を導き出してしまう可能性もある (Campbell & Kenny, 1999)。

このような理由からも，研究では必ず検定を行って結果を報告しなければならないというわけではなく，何よりも手元のデータの持つ意味をしっかりと吟味するという姿勢が大切である。また，この方法を取る場合には，研究者の行き過ぎた恣意性を排除するために，第三者の意見や，該当学習者の声を入れるなどの手順も考慮に入れるべきであろう。

5. おわりに

いくつかの従属変数に対して検定を繰り返すという，検定の多重性の問題については，説明がなされている統計手法の解説書があまりないためか，第 1 種

の誤りを犯していることを気づかずに結果を議論している研究論文が多くみられる。そこで，本研究ではそのような問題を検討するため，従属変数（テストや質問紙の項目）が 5 つある場合に *t* 検定を繰り返すことにより第 1 種の誤りを犯す確率を，モンテカルロ・シミュレーションを用いてどう変化するかを調査した。その結果，第 1 種の誤りを犯している確率（α）は，従属変数間にとても強い相関関係（$r = .95$）がある場合には低くなり，相関がない（$r = .00$）場合には高くなるが，すべての場合において有意水準として設定した 5% を超えてしまうということが明らかになった。ゆえに，複数の従属変数の検定でも多重性の問題への注意が必要であるということが確認されたと言える。

統計的な手法による研究結果の分析と解釈は，統計解析ソフトの普及により，利用される頻度が今後も増え続けるであろうと思われる。しかし，先行研究で用いられている誤った手法を，ブラック・ボックスの状態で理解しないまま使用するような慣習が続く限りは，以下のような批判を生む一因となるだろう。

> 過去 30 年ほどの日本の英語教育改革は，日本人が学校教育で思うように英語力が身に付かないのはもっぱら教え方が悪いせいであるとの間違った認識に基づいて進められてきた。そのため，英語教育関係者の関心が教授法に集中するようになってきた。そして，たかだか 2 ~ 30 人の学習者相手に数週間から数ヶ月ほどの実験授業を行い，その効果の測定結果を統計処理して有意差が出たのでないのと論じる疑似科学的教授法研究が横行するようになったのである。昨今，熱心な英語教師ほどその熱意を教授法研究に注ぎ込んでしまうのは，まことに残念なことだと言わざるを得ない。(斎藤, 2008, p. 41)

このような考えによって，「体系的な研究（学問）分野としての外国語教育学」自体が否定されるようなことがないように，また，確立された分野として学術的に進歩していくためにも，厳密な研究手法，そして，その分野での標準化されたルール[10]を研究者が身につける努力を続けていかなければならない。

謝辞

本稿を執筆するにあたり，常磐大学の小泉利恵先生，豊橋技術科学大学の印南 洋先生，広島大学の前田啓朗先生に大変貴重なご意見をいただきました。ここに記して感謝いたします。

注

1 分散分析を行う際に，前提を確認するための等分散性の検定などは，検定の志向するところが異なるので，検定の繰り返しとは見なされない (足立, 1998)。

2 実験群と統制群の能力が等しいということを証明するために，「有意差がない」($p > .05$ である)ことを論拠としている研究論文もあるが，有意差検定の考え方では，「有意差がない」=「等しい」ではないため，これは正しい方法であるとは言えない。ゆえに，4.3 節で説明している効果量や，信頼区間の報告をするのが望ましい (森, 2008)。しかし，このような正しくない検定の利用は SEM（Structural Equation Modeling）の χ^2 検定にも見られるように，一般的に使われていることが多い。

3 あるいは研究者が自分の研究目的に応じて，適切な値に定める。

4 データ解析環境 R とは，統計解析とグラフ作成を行うことができるフリーソフトであり，The R Project for Statistical Computing (www.r-project.org) よりダウンロードが可能。

5 モンテカルロ（Monte Carlo）は，地中海に面するモナコ公国の都市の名前である。モンテカルロはカジノで有名なリゾート地であり，乱数でデータを発生させる様子が，サイコロを何度も振る様子を連想させることから，モンテカルロ・シミュレーションという名前がつけられたと言われている。

6 同様のシミュレーション研究が，Hummel and Sligo (1971) をはじめとして，かなり以前から行われているが，今回の研究は複数の従属変数における検定の多重性の問題に焦点を当てている点で異なる。

7 効果量 d の大きさの目安は，.20（効果量小）.50（効果量中）.80（効果量大）となっている (Cohen, 1988)。

8 同様の定義として familywise error rate があり，これらの違いは Weinfurt (1995) を参照。

9 0.01+0.01+0.01+0.01+0.01=0.05 となっているので，α が .05 を超えていない。

10 ここでの「標準化されたルール」とは，APA Publication Manual などの，各分野で定められているルールにきちんと従って論文を執筆することも含まれている。なぜならば，「ルールを作り上げて，その土俵の上で質を競うというのが，プロフェッショナルの世界の掟であり，標準化されたルールがない，もしくはルールを守らないということは，その研究分野が成熟していないことを意味する」(竹内 理, 私信, 2008 年 2 月 3 日) ためである。

参考文献

足立堅一 (1998). 『らくらく生物統計学』 東京：中山書店.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: American Psychological Association.

Bray, J. H., & Maxwell, S. E. (1982). Analyzing and interpreting significant MANOVAs. *Review of Educational Research*, *52*, 340–367.

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York: The Guilford Press.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. Retrieved September 21, 2007, fromhttp://www.psycho.uni-duesseldorf.de/aap/projects/gpower/

Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). London: SAGE.

Field, A., & Hole, G. (2003). *How to design and report experiments*. London: SAGE.

舟尾暢夫 (2005). 『The R tips』 東京：九天社.

南風原朝和 (2002). 『心理統計学の基礎－統合的理解のために』 東京：有斐閣.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Huberty, C. J., & Morris, J. D. (1989). Multivariate analysis versus mutiple univariate analyses. *Psychological Bulletin*, *105*, 302–308.

Hummel, T. J., & Sligo, J. R. (1971). Empirical comparison of univariate and multivariate analysis of variance procedures. *Psychological Bulletin*, *76*, 49–57.

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*, 219–244.

石井秀宗 (2005). 『統計分析のここが知りたい』 東京：文光堂.

磯田貴道 (2004). 「テストの結果を比べる: 3 クラス以上の場合」 前田啓朗・山森光陽 (編著) 磯田貴道・廣森友人 (著) 『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』 (pp. 42–52). 東京：大修館書店.

磯田貴道 (2008). 『授業への反応を通して捉える英語学習者の動機づけ』 広島：溪水社.

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

Koizumi, R., & Katagiri, K. (2007). Changes in speaking performance of Japanese high school students: The case of an English course at a SELHi. *ARELE*, *18*, 81–90.

栗田佳代子 (1999). 「実際のデータを用いた t 検定および検定力分析の「観測値の独立性」からの逸脱に対する頑健性の検討－人工データによる研究結果との対応および項目の尺度化の影響を中心に－」 『教育心理学研究』 *47*, 263–272.

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－」 *JLTA Journal*, *6*, 140–147.

前田啓朗 (2008a). 『多変量解析について』 Retrieved September 8, 2008, from http://home.hiroshima-u.ac.jp/keiroh/maeda/statsfaq/manova.html

前田啓朗 (2008b). 「WBT を援用した授業で成功した学習者・成功しなかった学習者」 *ARELE*, *19*, 253–262.

Meara, P. (2005). Lexical Frequency Profiles: A Monte Carlo analysis. *Applied Linguistics*, *26*, 32–47.

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.

森 敏昭 (2008). 「Q16 検定のロジック」 繁桝算男・柳井晴夫・森 敏昭 (編著) 『Q & A で知る統計データ解析 DOs and DON'T's [第 2 版]』 (pp. 30–33). 東京：サイエンス社.

村井潤一郎 (2006). 「サンプルサイズに関する一考察」 吉田寿夫 (編著) 『心理学研究法の新しいかたち』 (pp. 114–141). 東京：誠信書房.

永田 靖・吉田道弘 (1997). 『統計的多重比較法の基礎』 東京：サイエンティスト社.

小野寺孝義 ・ 菱村 豊 (2005). 『文科系学生のための新統計学』 京都：ナカニシヤ出版.

斎藤兆史 (2008). 「英語教師は, まず優れた英語の使い手たれ」『英語教育』 *57*(7), 41.

芝 祐順・南風原朝和 (1990). 『行動科学における統計解析法』 東京：東京大学出版.

靜 哲人 (2007). 『基礎から深く理解するラッシュモデリング －項目応答理論とは似て非なる測定のパラダイム－』 大阪：関西大学出版部.

靜 哲人・竹内 理・吉澤清美 (2002). 『外国語教育リサーチとテスティングの基礎概念』 大阪：関西大学出版部.

Tabachnick, B. G., & Fidell, L. S. (2006). *Using multivariate statistics* (5th international ed.). Boston, MA: Pearson/Allyn & Bacon.

竹原卓真 (2007).『SPSS のススメ―2 要因の分散分析をすべてカバー』 京都：北大路書房.

Weinfurt, K. P. (1995). Multivariate analysis of variance. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 245–276). Washington, DC: American Psychological Association.

山田剛史・杉澤武俊・村井潤一郎 (2008). 『R によるやさしい統計学』 東京：オーム社.

Appendix A　シミュレーションで使用した R のコマンド例

```
montecarlo <- function(n) {                          # 関数の定義
count <- 0                                           # カウンタを 0 に戻す
for (i in 1:n) {                                     # 処理を n 回繰り返す
group<-c(rep("A",100),rep("B",100))                  # A と B のグループを各 100 作る
factor<-factor(group)                                # 要因型ベクトルに変換
testA1<-rnorm(n=100,mean=50,sd=10)                   # 正規分布から標本を抽出（testA1）
testA2<-rnorm(n=100,mean=50,sd=10)                   # 正規分布から標本を抽出（testA2）
test1<-c(testA1,testA2)                              # test 1 としてまとめる
result1<-t.test(test1~factor,var.equal=TRUE)         # t 検定の結果を result 1 にまとめる
testB1<-rnorm(n=100,mean=50,sd=10)                   # 以下，result 5 まで同じ作業
testB2<-rnorm(n=100,mean=50,sd=10)
test2<-c(testB1,testB2)
result2<-t.test(test2~factor,var.equal=TRUE)
testC1<-rnorm(n=100,mean=50,sd=10)
testC2<-rnorm(n=100,mean=50,sd=10)
test3<-c(testC1,testC2)
result3<-t.test(test3~factor,var.equal=TRUE)
testD1<-rnorm(n=100,mean=50,sd=10)
testD2<-rnorm(n=100,mean=50,sd=10)
test4<-c(testD1,testD2)
result4<-t.test(test4~factor,var.equal=TRUE)
testE1<-rnorm(n=100,mean=50,sd=10)
testE2<-rnorm(n=100,mean=50,sd=10)
test5<-c(testE1,testE2)
result5<-t.test(test5~factor,var.equal=TRUE)
count <- count + ifelse(result1$p.value<0.05 | result2$p.value<0.05 |
result3$p.value<0.05 | result4$p.value<0.05 | result5$p.value<0.05, 1, 0)
                                                     # 第 1 種の誤りの判定
}
return (count/n)                                     # 結果を出力
}
```

```
montecarlo(繰り返しを行う回数)                        # 関数を指定した回数だけ実施する
```

Note. 変数間に相関を持たせる場合のコマンドは，山田・杉澤・村井 (2008, p. 326) を参照に作成したものを用いた。

Appendix B　シミュレーションで発生させた 5 つの従属変数間の相関係数の例

(a) r = .00（無相関）

	Test 1	Test 2	Test 3	Test 4	Test 5
Test 1	—				
Test 2	-.10	—			
Test 3	.09	.02	—		
Test 4	.11	.07	.03	—	
Test 5	.08	-.06	-.04	.03	—

(b) r = .50（中程度の相関）

	Test 1	Test 2	Test 3	Test 4	Test 5
Test 1	—				
Test 2	.49	—			
Test 3	.50	.48	—		
Test 4	.45	.44	.44	—	
Test 5	.53	.48	.52	.51	—

(c) r =.95（とても強い相関）

	Test 1	Test 2	Test 3	Test 4	Test 5
Test 1	—				
Test 2	.95	—			
Test 3	.95	.95	—		
Test 4	.96	.96	.95	—	
Test 5	.96	.95	.95	.95	—

Note. 乱数を使って無作為抽出を行っているので，毎回近い値は得られるが同じ値にはならない。