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 2014 年度 第 6 号報告論集
草薙 邦広（pp. 46–84）

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―

草薙 邦広
名古屋大学大学院
日本学術振興会

概要

外国語教育研究におけるデータ分析の手法は，2000 年代よりめまぐるしいほど高度化している。しかしながら，日頃の教育実践のなかで，学術的知見や，それを支える高度なデータ分析の結果を適切に理解し，そして自らの意思決定に役立てることは容易ではない。たとえば，メタ分析などでもちいられる効果量（effect size）の根本的概念は，外国語教育に携わるものの実務的観点からみて，けっして親和性の高いものではない。そこで，本稿では，はじめに，実験計画法，統計的仮説検定，そして効果量とその信頼区間の算出といった方法について紹介し，これらが，集団に対する処遇の結果を解釈するという文脈における実務的な観点と，やや乖離しているいくつかの点（e.g., 解釈困難性，中心傾向への依存）について示す。つぎに，効果量を，より解釈が容易なかたちに変換した指標である効果偏差値（e.g., 伊藤, 1998）および優越率（e.g., 南風原, 2014; 南風原・芝, 1987）を紹介する。最後に，効果量のみでは解釈できない部分を補うための，いくつかのあたらしい定量的方法（e.g., 比較点，分位点回帰をもちいた分析法）を提案する。

Keywords: 指導法，研究方法論，統計，効果量，教育評価

1. 背景

ある特定のカリキュラム，教育プログラム，指導法，教室内外の活動，そして学習者が自律的にとる学習方法などが，いったいどれほどの効果をもつものなのか，という観点は，非常に素朴ながら，教育従事者にとって常に最大の関心事のひとつである。当然ながら，教育従事者は，日頃の教育実践において，無数の意思決定に迫られている。しかし，概して，その意思決定の「正しさ」に対する担保や客観的な証拠は，容易に得られるものではない。たとえば，今学期，英語科で一丸となって取り組んでいるあたらしい教室内活動や生徒に課している課題が，従来のものとくらべ，成績にどのように影響しているのか，または，昨年度に一新した新カリキュラムの効果はいかほどであるか，というような観点は，いうまでもなく，教育実践に直接的に関わる問題である。しかし，それを

客観的に測定・評価するためには，第一に，人的な資源が必要である。教育従事者の大部分にとって，通常の業務にくわえ，こうしたデータ分析をおこなうことは，たやすいことではない。また，定量的な方法による集団に対する処遇の結果（treatment outcome）の解釈には，言語テスト，統計学，そして教育評価などに関わる多少の専門性を要するものであるし，定性的な観察は，むしろ，それ以上の訓練を要する。

教育従事者のなかに，ある処遇の結果を検討するためのコストが高い，このこと自体を好ましいととらえるものは，いないであろう。しかし，一方で，外国語教育に関わるデータ分析の手法は，飛躍的に高度化してきている。特に近年は，統計的仮説検定に依拠していた従来の分析方法にくらべ，全体的に，分析上の制約を減らし，結果が頑健で，より自由な手法をよしとするようになってきている（Larson-Hall, 2012; Larson-Hall & Herrignton, 2010）。なかでも，効果量やその信頼区間（後述）を重視するという現在の顕著な風潮は，従来のスタンダードであった統計的仮説検定に対する依存の脱却をめざしているという点で，まさに「統計改革」ともいえるであろうし（e.g., 大久保, 2009），体系的に過去の研究成果を統合するメタ分析によって，外国語教育研究は，円熟期に達したかのようにもみえる。しかし，効果量やその信頼区間の重視といった動き，そして，それ以外のさまざまなデータ分析技術の向上が，外国語教育に関する実務にどれほど還元されているかといえば，いまだにいくばくかの乖離があるといわざるをえない。

また，同じような現象を対象とする場合でも，実務的ないし実用主義的な観点と，学術的ないし理論実証主義的な観点では，ものごとの見方やアプローチが異なるというのは，世の常である。このことは，外国語教育においても例外的でないようにおもわれる。たとえば，本稿の大部分を割いて，後にくわしく説明するが，外国語教育研究や第二言語習得研究におけるメタ分析（e.g., Norris & Ortega, 2006）でもちいられる，統計的な効果を示す指標（効果量，effect size）の一部は，外国語教育に関わる，いくつかの実務的な観点のもとでは，かならずしも解釈しやすいものではない。これは，実務的な観点と学術的な観点が，根本的なところで異なることに由来するかもしれない。

そもそも，学術的な観点のもとでは，効果量自体に対して実質科学的な解釈をすることは，かならずしも必要なことではない。むしろ，学術的な観点では，対象とする現象の効果を，十分な精度ないし確証をもって観測し，その観測に対して整合的な理論体系を築くことを重要視する。また，効果量は，種々のモデリングや統計的仮説検定の前提となる諸概念の根幹をなすものである。

一方，実務的な観点では，効果を実質科学的に解釈し，直面する問題の解決や意思決定に寄与させていくことを重要視する（e.g., 南風原, 2014）。ある処遇の効果が大きいか，小さいか，といった解釈は，処遇のありかたを選択するための直接的な材料になりうるし，さらに，見込まれる結果に対する予測の材料にもなりうるとされている。

しかしながら，学術的な観点のもとではなく，教育従事者の意思決定のなかで，効

果量による処遇の結果の解釈や，メタ分析による知見が，有効に活用「されているか」，そして，その前に，そもそも活用「されうるのか」という観点が明示的に論じられているわけではない。「効果量の値の解釈は，実質科学的な結果の解釈に役立つ」という一般的な見方は，外国語教育において典型的な，集団に対する処遇の結果を解釈する場面においても，はたして正しいのだろうか。本稿は，まさに，この効果量の値の解釈を論の軸としながら，外国語教育に関する実務的観点のもとで，集団に対する処遇の結果を，定量的方法をもって解釈することについて考察するものである。

本稿では，最初に，集団に対する処遇の結果を検証するためのデータ収集方法（準実験計画法）について，つぎに，統計的仮説検定と，効果量およびその信頼区間について，簡便に概論をのべる。その後，効果量の一部に関連する概念が，さまざまな面において，外国語教育に関する実務的な観点にそぐわない点があることをのべる。つぎに，これらの相違点を補うであろう，いくつかの方策（e.g., 効果偏差値，優越率，比較点，分位点回帰をもちいた分析）について紹介する。

本稿のなかには，一部，数学的および統計的な知識を要するようにみえる部分がある。しかし，このことによって，読者の関心を削いだり，読者の理解を困難にすることは，著者の本意ではない。そうした部分を読み飛ばしても，できるだけ，論が通るようにこころがけたつもりである。

2. 前提

2.1 準実験計画法と統計的仮説検定

集団に対する処遇の結果を検討するためのデータ収集方法には，実にさまざなものがある。外国語教育研究および第二言語習得研究では，実験協力者の無作為な標本化（random sampling），無作為なグループへの割りあて（random asignment），そして実験者による変数の完全な操作が困難であるため，このような条件を満たさない準実験計画法（quasi-experimental design）とよばれる手法をもちいることが多い。

準実験計画法のなかでも，実験計画はいくつかにわかれる。外国語教育研究において，もっとも頻繁にもちいられるものは，非同一集団計画（non-equivalent group design, NEGD）である。NEGD では，実験協力者を，複数の群（実験群と統制群ないし処置群と比較群）に対して無作為方式ではない方法で割りあて，処遇を挟んで試験（検査，実験）をおこない，成績を比較する。外国語教育では通常，学年，学級，受講している授業などによって各群への割りあてをおこなうため，このような場合は無作為方式ではないといえる。

そのほかにも，非同一変数計画（non-equivalent variable design，NEVD）とよばれる方法がある。NEVD は，NEGD と異なり，単一の群を対象とする。しかし，NEGD と同様に，単一の群（実験群）に対して，処遇を挟む 2 点間の試験をおこなう。この試験では，それぞれ複数の変数（テストなど）を計画に入れる。たとえば，暗記によって語彙学習を

おこなう処遇をある集団に対して実施したとして，事前と事後に，それぞれ語彙テストと読解テストをおこなう，という具合である。この場合，群間ではなく，変数間の比較によって効果を検証する。

ここでは名前をあげるだけに留めるが，準実験計画法には，これらのほかにも，さまざまな手法がある。事前テストの成績において，任意の閾値以下の成績をもつ実験参加者のみに対して処遇を実施し，事前―事後における回帰直線の差分を検討する不連続回帰デザイン（regression-discontinuity design，RDD），NEGD のうち，事後のデータのみを収集する計画（only post-test design），さらに，単一事例研究（single case study）などがある。本稿では，これ以後，準実験計画法のうち，もっとも典型的であると考えられる NEGD で得られるデータを議論の基本軸としていく。

準実験計画法によって収集したデータは，通常，一般線形モデル（e.g., 分散分析，共分散分析，多変量分散分析，多変量共分散分析）をもちいて分析する。NEGD のデータに対して分散分析をもちいる場合では，独立変数を群（実験群と統制群の 2 水準，被験者間要因）および時期（事前と事後の 2 水準，被験者内要因）とし，従属変数を試験の成績とした混合計画をもちいる。このとき，ふたつの独立変数の交互作用が統計的に有意であれば，処遇がそれぞれの集団における平均値の変動におよぼす影響が一様ではない，ということなのだから，処遇になんらかの効果があったと解釈できる。

統計的仮説検定（statistical hypothesis testing）の観点において，「統計的に有意である」ということが示すのは，「母平均差が 0 である」といった統計的帰無仮説（null hypothesis, H_0）の前提のもとで，観察された統計量に対する確率論的な整合性が低い，ということである。よって，統計的有意性のみをもって，処遇の結果を適切に解釈できる場合は，非常に限られている。簡便な例をあげると，スチューデントの t 検定をもちいる平均差の検討における有意確率は，検定統計量 t と自由度に対応する確率である。検定統計量である t は，効果の大きさ（くわしくは後述）と推定の精度（標本の大きさ）の積であるといえる（e.g., 南風原, 2002）。そのため，統計的有意性が得られたとしても，効果が小さい場合がありうるし，逆に統計的有意性が得られなくとも，効果が大きい場合もある（e.g., Kline, 2004; 水本・竹内, 2008）。つまり，統計的有意性のみをもって，効果の大きさを議論することは基本的にはできない。さらに，処遇の結果が統計的有意性を示したとしても，かならずしも，それが教育実践上，望ましい処遇の担保になるとはかぎらない。

統計的仮説検定の考えかたのもとでは，むしろ，適切に統計的仮説検定をおこなうために，前もって予測される効果の大きさに対して適切な標本サイズ（N）を設定する必要がある。このためには，検定力分析（power analysis）とよばれる手続きをおこなうとよい（e.g., Erdfelder, Faul & Buchner, 1996; 水本・竹内, 2011; 豊田, 2009）。そもそも，学術的な観点のもとでは，研究の対象とする効果の大きさに，原理的な制限があるわけではない。たとえば，研究の対象が，微小の効果しかもたない現象であっても，逆に大きな効果をも

つものでも，その一切は，学問的な文脈および研究者の関心に依存する。微小の効果のみをもつ現象を，ある程度の精度をもって観測するためには，必然的に大きな標本が必要となる。その逆に，大きな効果を観測するときには，小標本であっても十分な検定力が得られる場合がある（e.g., 永田, 2003）。しかし，教育実践に関わるかぎり，標本サイズを適切に統制することは，容易ではない（e.g., 草薙, 2014a）。

よって，データ収集の方策としての準実験計画法の適切さは別として，統計的仮説検定の有意性という観点は，処遇の結果を解釈する道具立てとして，かならずしも完璧なものではない。もちろん，この言明は，一般的な統計的仮説検定の是非を論じるものではないし，処遇の結果を解釈する場面における，統計的仮説検定の有効性のすべてを疑うものでもない。

2.2 単純効果量と標準化効果量

本稿では，これまでさまざまな文脈において，「効果」（effect）ということばをもちいてきた。しかし，このことばには，さまざまな意味がある。一般的に，統計学における効果とは，データがもつ全体のばらつきにおける，任意の（研究者が関心をもつ）一部分を示す（e.g., Cohen, 1988）。たとえば，2 群のデータには，それぞれの群内における固有のばらつきがある。群内のばらつき（散布度）を数量化するためには，分散，標準偏差や四分位区間などをもちいる（南風原, 2002）。こうしたばらつきがあるなかで，群間によるばらつきは，一種の効果とみなすことができる。また，2 変量のばらつきをあらわすためには，共分散をもちいる。相関係数は，共分散を 2 変量それぞれの標準偏差で割ったものでもあるのだから，相関係数も一種の効果であるとみなせる。

しかし，このような概念を，実質科学的な意味における効果と，容易に置き換えてはならない。たとえば，ある処遇の結果の望ましさと，上記のような統計的な意味における「任意のばらつき」は，同一視できないときがある。処遇の結果の望ましさは，実務的な文脈や目的によって，そのありさまが多岐にわたる。一方で，統計的な効果は，そのような文脈や目的を数量的に反映するわけではない。

統計的な効果を数量化するためには，効果量とよばれる指標をもちいる（和文の平易な概説としては，大久保・岡田, 2012; 水本・竹内, 2008,）。効果量の指標にも，もちろん，さまざまなものがある。まず，単純効果量（非標準化効果量）と標準化効果量を区別するべきである（e.g., Frick, 1999; Olejnik & Algina, 2000）。

典型的な場合，1 変数 2 群ないし 2 変量 1 群の単純効果量は平均差（mean difference, md）である。NEGD における実験群と統制群の比較を念頭におくと，実験群と統制群におけるそれぞれの平均値（μ）をもちいて，

$$\mathrm{md} = \mu_2 - \mu_1$$

とあらわせる。ただし，添字は群をあらわし，1 を統制群，2 を実験群としている。

当然ながら，平均差は測定具のスケール上であらわされている。そのため，この値をもって，実験群のほうが 20 点成績が高い，というような解釈ができる。しかし，異なるスケールをもつもの，つまり，測定が異なるもののあいだでは，平均差を直接比較することはできない。たとえば，990 点満点の TOEIC のスコアと，10 点満点の語彙テストのスコアを比較しても基本的には無意味である。

そのような場合，測定具のスケールに依存しない標準化効果量をもちいる。標準化効果量にも，実にさまざまな種類があるが，ここでは，主として，標準化平均差（standardized mean difference）を取りあげる。標準化平均差は，*d* 族の効果量とよばれるものであり（e.g., Cohen, 1988; 大久保・岡田, 2011; 水本・竹内, 2008, 2011），Cohen's *d*, Hedge's *g* や Glass' ⊿といった指標がある。まずは，NEGD のデータを分析する際において，最も一般的な指標である，Glass' ⊿の定義を見てみる。⊿は，2 群の平均値（μ）と統制群の標準偏差（σ）から，以下のようにもとめられる。

$$\Delta = \frac{\mu_2 - \mu_1}{\sigma_1}$$

⊿などの標準化平均差は，得点のばらつき自体を単位（分母）とし，集団を基準として標準化されているため，異なる測定単位間の効果を比較することができる。このことから，研究間で測定単位が違えども，ふたつの処遇の標準化平均差が 1.00 であれば，平均差と標準偏差の大きさが，両方で同じ程度である，というように解釈できる。複数の研究成果を体系的に統合するメタ分析（外国語教育における代表的な研究として，Norris & Ortega, 2006 がある）は，このような測定の標準化によって，はじめて可能になる。しかし，外国語教育研究で，異なる測定単位（テスト方式）による処遇の結果を，直接的に統合したり，比較したりすることには一定の批判もある（e.g., 亘理, 2014）。

近年，外国語教育研究および第二言語習得研究では，メタ分析が隆盛であり，非常に多様なテーマにおける研究事例がこれまで刊行されている。一般的な，群間の平均値の比較を取りあつかうメタ分析では，標準化効果量として，⊿よりも Cohen's *d* をもちいる場合が多い。参考までに，*d* を算出する方法についても触れる。計算は⊿よりもやや複雑で，まず，群間でプールされた標準偏差（pooled standard deviation）をもとめる。*n* を標本サイズ，*σ* を各群の標準偏差，下の添字を群とすると，

$$\sigma_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^1}{n_1 + n_2 - 2}}$$

となる。これは，2 群の標本サイズが等しければ，

$$\sigma_{\text{pooled}} = \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}$$

としてもよい。⊿と同様に，標準偏差で平均差を割り，

$$\text{Cohen's } d = \frac{\mu_2 - \mu_1}{\sigma_{\text{pooled}}}$$

としたものが *d* である。

標準化平均差は，上記のように，測定単位によらず，さらに統計的有意性という二値的判断に終始する統計的仮説検定の枠組みとは異なり，その値の大小自体を実質科学的に解釈することができるとされている（e.g., 大久保・岡田, 2012）。そのため，国内外を問わず，さまざまな研究分野で効果量報告の重要性が主張されている（e.g., Kline, 2004; 大久保, 2009; 大久保・岡田, 2012）。

2.3 効果量の誤差

しかしながら，通常，手元のデータから得られた効果量の値は，標本効果量にすぎない（e.g., 南風原, 2002）。効果量にも標本誤差がある。仮に，同一条件の実験計画を繰り返したとして，その度ごとに，標本効果量が同じ値をとるとはかぎらないし，処遇の結果における母効果量を知ることができるか，という点は論理的にも難しい問題である。

あらゆる要因に条件付けられた，過去のある処遇を，一度きりのできごとであったと考え，あくまでも，その処遇というできごとの再現性はなく，議論の対象は，処遇を受けたひとのみである，と考えることもできる。この場合，得られた効果量を母効果量ととらえても差し支えない。一方，特定の特性をもつ処遇を，特定の特性をもつひとに実施した場合，というように，ひとと処遇の組み合わせについて一般化した母集団を想定すると，得られた効果量は標本効果量にすぎないともいえる。これは，研究者や教育実践者のそれぞれが，ときと場合に応じて考えるべき問題であるが，ここでは，便宜的に後者の考えかたをとる。

さて，得られた標本効果量には誤差があるとして，その誤差の大きさは，標本サイズに規定されている。このことは，たとえば，*d* における誤差（*SE*）の計算式が，

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

であることからもわかる（e.g., 南風原, 2002）。

標準誤差をもちいると，任意の幅の信頼区間（confidence interval，CI）を推定することができる。信頼区間とは，母数に対して点ではなく区間で推定したものである。一般的に広くもちいられるものは，95%信頼区間である。標準誤差から，95%信頼区間の下限値と上限値をもとめるためには，

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

とする。

外国語教育研究において一般的であろう，40 人程度（学級規模）の比較では，d = 0.50 として，その 95%信頼区間の下限値が 0.05，上限値は 0.95 程度である。これは，研究者の直感よりも広いものとしてとらえられるかもしれない（e.g., 豊田, 2009）。この点からも，標本効果量の値を，非常に固定的なものとしてとらえるべきではないことがわかる。さらに，区間推定の下限値や上限値も，標本値をもちいて推測した値にほかならないのであるから，その点も注意すべきである（e.g., 南風原, 2014）。

よって，統計的有意性が得られないときに，効果量の値のみをもって議論を積極的に進めることは，勇み足になりかねない。統計的有意性がないということは，その実験で設定された標本サイズでは，対象とする効果を適切な精度をもって観測できていない，ということにほかならない。つまり，推定された効果量の値自体，さほど信用できないということである（e.g., 水本・竹内, 2011）。効果量の値を実質科学的に解釈することが重要であるとされているが，その効果量の値の確からしさも，ときに，それ以上に重要な観点であるといえる。

ちなみに，対応がない比較における単純効果量の標準誤差は，以下のようにしてもとめる。

$$SE_{md} = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{(n_1 - n_2 - 2)}} \times \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

この値は，群間でプールされた標準偏差に対して，標本の大きさで重みづけをしたものとおなじである。平均差を，この値で割ったものが，対応がない 2 群の平均差のための検定における検定統計量 t となる。

$$t = \frac{(\mu_2 - \mu_1)}{SE_{md}}$$

標本効果量の報告だけでなく，その信頼区間の推定は，現状では，研究を公刊するうえで，非常に優れた方策とされている（e.g., 大久保・岡田, 2012）。また，実験計画の際に，目標とする効果量の信頼区間を満たす標本サイズを，前もって決定することができる（e.g., 南風原, 2002）。効果量などにもとづく標本サイズ設計を正確度分析（precision analysis）とよぶ。

3. 集団に対する処遇の結果を解釈する際の効果量

このように，効果量は，その誤差にさえ注意すれば，集団に対する処遇の結果をあらわす指標のひとつになりうる。たとえば，統制群と実験群の平均差や標準化平均差は，処遇の効果とみなせる場合もある。平均差や標準化平均差の値が大きければ，処遇が成績におよぼす影響が強いといえるため，平均差や標準化平均差が大きいことが既知である処遇を，教育実践に積極的に取り入れてもよいかもしれない。

Norris and Ortega（2006）をはじめとして，外国語教育研究では，実にさまざまな処遇の結果についてのメタ分析が刊行されている。また，近年の，教育実践の報告においても，集団に対する処遇の結果をあらわすものとして，効果量がもちいられるようになってきている。これらの試みが，われわれに，重要な知見をあたえていることは，疑いようのないことである。

しかし，非常に限られた場面ではあるが，教育実践を念頭に置いたとき，集団に対する処遇の結果をあらわすために効果量のみを利用することが，常に適切である，とはいいがたいかもしれない。本稿における目的の一部は，効果量の利用，それ自体に関わるものではなく，この限られた文脈における，効果量の利用に関する実務的事情について，多少の考察をすることである。

まず，結論からのべると，著者の主張は，効果量の利用は，外国語教育研究や第二言語習得などの学術的な立場ではまだしも，教育実践やそれに関わる実務的ないくつかの観点のもとで親和性に欠く面もある，というものである。ここで，著者が，教育実践や，それに関わる実務的な観点とのべるのは，特定の処遇，つまり，あるカリキュラム，教育的プログラム，指導法，教室内外の活動，学習方法について，なんらかのデータにもとづいて，意思決定をする，という場面を念頭に置いている。この想定自体，一般的なものではないかもしれない。しかし，仮に，そのような場面における効果量の利用について考察することによって，一般的な文脈における効果量の利用についての理解が，より深まるかもしれない。

以下には，集団に対する処遇の結果を解釈する際に，効果量を利用することが，ど

のような点において，実務的観点との親和性に欠けるか，具体的な 8 つの点をあげている。これらのほとんどが，本質的には，(a) 値自体の解釈が困難であること，(b) 効果量が分布の中心傾向のみをあらわすこと，このふたつに由来するため，それぞれの点が，複雑に関わりあっていることを，前もってのべておく。

3.1 標準化効果量の値の解釈が容易ではないこと

効果量は，実質科学的な解釈に役立つとされているものの，標準化効果量（標準化平均差など）は，前述のとおり，測定のスケールに依存しない。しかし，逆に測定のスケールに依存しない値だからこそ，その効果をイメージすることが困難になってしまう（e.g., 伊藤, 1998）。たとえば，$d = 0.20$ が，実質科学的にどれほどの効果であるかは，けっして理解しやすいものではない。秋田県在住と愛知県在住の男子高校生の身長について標準化平均差をもとめると，およそ $d = 0.20$ となる。しかしこの値自体を，直接的に，そして直感的に理解することは，ほとんどのひとにとって不可能である。

後述するが，効果量は，標準正規分布 N（0 , 1）のスケールであらわされており，これは標準化得点ないし z 得点ともよばれる（図 1）。標準化得点のスケールに対して，親しみをもつものならまだしも，そうでないものにとって，効果量の値自体は親和性の高いものにはなりえないだろう。

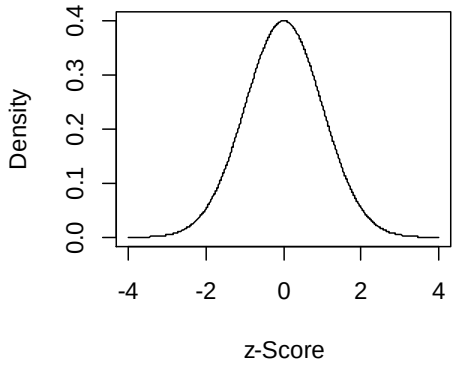


図 1. 標準正規分布 N（0 , 1）の確率密度曲線

しかし，この点に関しては，単純効果量による解釈が有効である。平均で 10 点の伸び，という実測値のスケールにおける情報は，きわめて直接的な理解をもたらす。

しかし，逆に，複数の処遇の結果を比較する際には，測定のスケールがすべて同一でなければならなくなってしまうし，ばらつきが大きいデータの平均差と，小さいデータの平均差を，直接的に比較しても不適切になるかもしれない。図 2 には，そのような場合を示している。

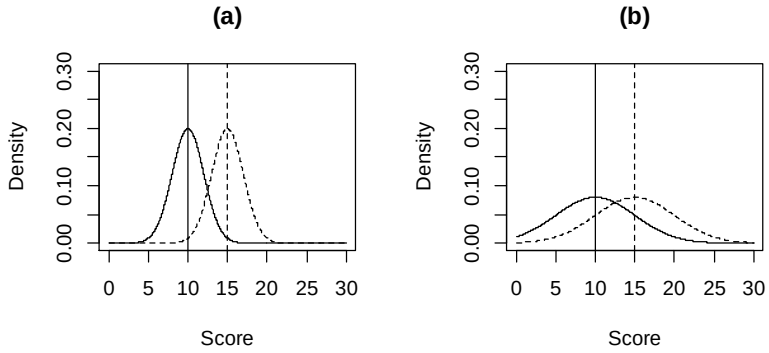


図 2. 同じ平均差をもつがばらつきが異なる 2 群の分布の例

図 2 の（a）および（b）は，ともに平均差が 5 であり，単純効果量としては同じ値をもつ。しかし，（a）のほうが，（b）よりもあきらかにばらつきが小さい。単純効果量のみでは，このような違いを区別できない。ほとんどの場合，ばらつきが大きければ大きいほど，誤差による変動も大きくなるため，ばらつきが大きく異なるものの平均差を比較しても有益ではない。

ちなみに，図 2 の（a）が $d = 2.50$，（b）が $d = 1.00$ となる。この場合は，Δも同値となる。

3.2 解釈基準が文脈に依存すること

このように，標準化効果量の値自体を，直接的に解釈することは容易でないため，解釈のための基準が使われることがある。一般に，Cohen（1988）による提案が基準として参照されることが多い。Cohen の基準は，d の場合には，0.20 を小，0.50 を中，そして 0.80 を大としている。Δなど，およそほとんどの標準化平均差はこの基準に準ずるであろう。しかし，もっとも重要な点であるが，Cohen 自身がのべているように（e.g., Cohen, 1998, 1994），このような解釈基準を固定的にとらえるべきではない。

効果量は，研究分野や，研究者が対象とする現象自体に強く依存するために，分野ごとに固有のベンチマークを開発しようという試みもある。外国語教育研究や第二言語習得では，Plonsky and Oswald（2014）が，過去の刊行論文を参照して，外国語教育研究のための，独自の効果量の解釈基準を提案している。このような考えかたは，学術論文の出版などに関するガイドラインとしての解釈基準が，似た現象を研究対象とするもののあいだで共有される，という点において，非常に有益であろう。

しかし，その本意については，浅学者の筆者が理解できるところではない。ある研究テーマの知見が蓄積されれば，対象とする効果量は小さくなっていくということが，一般的に広く知られている。外国語教育研究が，現在より発展すれば，将来，刊行されるであろう論文の効果量は，相対的に現在よりも小さくなっていくかもしれない。そうすれば，再度，効果量の解釈基準を，その地点までの刊行論文にあわせて変えるべきであろうか。

さらに，外国語教育研究は学際的であり，その測定具は，ほかの学術分野にくらべて非常に多様である。これは理論面においても，まったく同様である。認知科学に依拠する研究もあれば，社会心理学に依拠する研究もある。将来，このような多様性をうけて，外国語教育研究に属する，さまざまな下位の研究領域ごとに，同様のベンチマークを作成していく必要があるのだろうか。

むしろ，統計学的な意味での効果として，効果量が示すところは，平均差と標準偏差の比や，説明できる分散の比率であって，その意味が分野によって細かく変わるなどということはない。そして，「標準化平均差 1.00 は平均差と標準偏差の比率が等しいことを示す」といった統計的な事実は，各分野における過去の刊行論文に依拠することではない。いいかえれば，統計的な意味での効果量の値（連続量）を，カテゴリカル変数（大・中・小）として振りわけるのに，実質科学的な知見は，まったく独立であってよい。

そうではなくて，研究分野や，過去の刊行論文に依拠するのは，実質科学的な意味での効果である。そして，統計的な意味での効果と，実質科学的な意味での効果の「結びつき」も，研究分野といった大きい枠組みや，その歴史にあるものではないはずである。そのふたつの「結びつき」は，ケース・バイ・ケースであって，その都度，その文脈において，当事者が適切に判断すべきである。これは，メタ分析についても，およそ同様のことがいえる。メタ分析では，メタ分析をおこなうものの選択基準やコーディングが，研究の質を規定する。メタ分析の読者は，そこに厳しい目を向ける（e.g., 亘理, 2014）。

ここで，以下の例を考えてみる。外国語教育研究のメタ分析では，仲介変数として取りいれられることも多い「処遇の期間」についてである。ある処遇を 30 分間施した場合と，90 分×6 授業で施した場合の効果量を比較したとする。それ以外の要因による影響は，両者で等しいと仮定して，結果，前者が $d = 0.20$ で，後者が $d = 0.80$ という観測を得た。このとき，Cohen の基準に照らしあわせ，前者の効果量は小，後者のそれが大として，なにか実質科学的な解釈をすべきだろうか。統計的な意味での効果ならば，まったく差し支えない。平均差と標準偏差の比率において，後者がはるかに大きいということには変わりない。しかし，そもそも処遇の期間が長い後者のほうが，統計的な意味での効果が大きいというのは，実務家にとっては，まったく自明のことであって，それほど有益な観点ではない。こうした文脈を度外視して，大・中・小というような基準のみをもって，議論を進めるべきでない。しかし，仮に，それぞれの文脈において，can-do リストのような，なにかの実質科学的な現象の記述が，こうした統計的な意味での効果の基準に関連づけられたならば，このかぎりではないかもしれない。しかし，外国語教育において，そのような試みはみられない。

つまり，過去に刊行されている論文からみて，得られた値が相対的に小さいか，とか大きいか，という見方よりも，「効果量が取りうる値は，研究対象や文脈自体に複雑に依存していること」を理解することが重要である。効果量の値の解釈は，きわめて複雑で

あり，状況依存的である。そのために，実務的な観点のもとでは，常に解釈が容易とはいえない。このことは，以下の節においてもくわしく触れる。

3.3 カテゴリカルな解釈基準が意味を失う場合が多いこと

これまで何度も書いているように，効果量が取りうる値は現象自体に依存する。微小の効果量を対象とするときもあれば，大きい効果量を対象とするときもある。対象とする効果量の値が極端に小さいとき，そして大きい現象を対象とする場合，現行の大・中・小といった解釈の基準は，そのスケール上での意味を容易に失ってしまいかねない。

たとえば，微小の効果量のみをもつ現象を考えてみる。極端な例だが，60 分の処遇（仮に動機づけに関する大学教員の講演）が高校生の言語学習適性（language learning aptitude）におよぼす影響をみるとする。ほとんどの研究者や実務家は，この処遇の効果量はほぼ 0 である，と考えるだろう。仮に，この処遇の結果と，もうひとつの処遇（IQ テストの練習問題）の結果を比較するとして，こちらの処遇も大きい効果量は望めないだろう。このとき，両方の標準化平均差は，Cohen の解釈基準のもとで，せいぜいが小以下であると予測できる。これでは，どちらの処遇の効果も同程度であった（またはどちらも効果がなかった）ということになってしまう。値を直接的に比較して，相対的な効果の大小を論じることはできるが，「ごく短期間の処遇が，言語学習適性におよぼす影響」という文脈において，解釈基準における大・中・小は意味をもたない。いずれも小以下である。もちろん，対象とする効果量が，大きな値（e.g., > .80）をとる現象についても同様である。いずれも小，いずれも大となるような効果量のあいだの比較では，解釈基準の重要性が下がる。つまり，カテゴリカルな解釈のみでは，情報が損失してしまう場合がある。標準化平均差は理論上，絶対値で $0 < d < \infty$ の値を取り（Olejnik & Algina, 2000），興味がある効果量の値は文脈に依存する，ということを忘れてはならない。

3.4 教育評価および目標規準準拠テストとの乖離

教育評価は，個々に対しておこなわれるものであり，集団の代表値のみをもって評価を議論することは，中心的なことではない。しかし，処遇の結果などを，個々のみではなく集団という単位で把握すること（集団評価）もまた，有効である（e.g., 荻野, 1983）。そもそも，荻野によれば，教育評価は，（a）個々に対する絶対的評価，（b）教育単位である学級集団を基準にした個々の評価，そして（c）広い地域を基準として，そのなかの位置を示す標準検査，という三者の調和がなされることが重要であるという（荻野, 1983, p. 102）。

効果量は，当然ながら，集団の特性をあらわす値である。たとえば，Δは，処遇後の個々の伸びの代表値とみることができる。効果量による処遇の結果の解釈は，上記の三者のうち，（c）にもっとも近いと考えられる。しかし，いうまでもなく，効果量のみによる

議論では，（a）や（b）といった教育評価上の要素は満たされない。

また，教育評価のなかでは，目標規準準拠テストと集団基準準拠テストの 2 種類があるとされている。目標規準準拠テストでは，個々の成績に関する情報は，教育目標に照らしあわせて判断され，カリキュラム開発や指導の手だての材料として，診断的にもちいられる。目標規準準拠テストはカリキュラムとセットになっており，評価と指導は表裏一体の関係にある。個々の成績は，他者の成績から独立しているという点で，絶対評価といえる（e.g., 梶田, 2006）。

一方，集団基準準拠テストの目的は，個々を集団内で位置づけることである。個々の位置づけを知ることもまた，教育上の意味をもつのであるが（荻野, 1983），この観点のもとにおけるテストは，基本的に，集団ないし個人を正しく弁別することが望まれる（e.g., 梶田, 2006）。個々の成績は，集団内で相対化されているという点で，相対評価といえる。集団基準準拠テストでは，個々の成績を偏差値であらわすことが多い（e.g., 梶田, 2006; 荻野, 1983）。偏差値の詳細については，後述する。

外国語教育における実際の教育評価や，それに準じる実務のなかでは，この両方の観点が重要であり，どちらかの側面のみをもつ，という評価の場面のみではないはずである。むしろ，厳密に区別できるわけでなく，どちらでもあるような場面も少なくないと考えられる。

さて，ここで，効果量に話をもどす。効果量，とくにΔは，そもそも，集団がもつ値であるが，別集団を基準に標準化された値でもあるといえる。よって，集団基準準拠の側面のみをもっており，目標規準準拠の側面はもたない。効果量のみによる処遇の結果の解釈では，目標規準準拠の考えかたとは，そもそもの根本が異なる。

外国語教育における，具体的な例をだす。ある特定のカリキュラムにおいて，新出文法項目の定着について，集団に対する処遇の結果を検証したいとする。個々の定着をみるのはなく，ここでは，集団としての定着をみたい。事前―事後でその文法項目に関する問題の成績を比較したとして，それが，仮に$\Delta = 1.00$ であったとする。このとき，この値をもって，教育目標や指導の観点が，集団として満たされたといえるかどうかは，まったくわからない。いいかえると，定着したかどうか，という観点を，集団間の距離から判断することはむずかしいのである。あえて，効果量のみをもって解釈しようとすると，究極的には，目標規準準拠テストでは，全員未知の項目が全員既知になることが一種の理想なのだから，対象とする効果量の値はきわめて大きい，と予想してもいいだろう。ただ，ここで，目標規準準拠テストには正規分布といった分布のかたちに対する制約がない，ということにも，注意が必要である（e.g., 梶田, 2006）。一方，効果量は，基本的に正規分布を前提としているという点にも乖離がある。しかしながら，こういった事情もさまざまな文脈に依存するということは申し添えておきたい。

一方で，対象とする現象によっては，効果量が有効である場合も多い。たとえば，

外国語の熟達度に関するもの，その構成技能（component skills），外国語学習や使用に関する心理特性（または関心・意欲・態度）などは，基本的には，集団で標準化したスケール上の変化を論じるしかないのであるから，効果量による解釈は適切になりえよう。このような構成概念（construct）の値の変化は，ものにもよるが，目標規準準拠テストの例とくらべると，比較的幅が小さいと考えられるため，効果量は相対的に小さいと予測できる。

問題は，前述したとおり，外国語教育の実務上，これらのどちらでもあるようなケースが多いということである。校種や学年の差など，さまざまな場合によって，様態は大きく異なりうる。効果量による理解のみにこだわるのではなく，その場に応じた適切な方策の選択をこころがけたい。しかし，目標規準準拠テストの側面が強い場合，効果量による処遇の結果の解釈は不適切であろう。

3.5 ばらつきが考慮されないこと

ここでは，特に⊿について取りあげるが，ほかの標準化平均差についても，部分的には，同様のことがあてはまる。[1]処遇の結果を解釈する際に，効果量のみを利用する方法では，処遇の結果として生じた実験群のばらつきが考慮されなくなってしまう。たとえば，効果量が大（ここでは，便宜的ないい方にすぎないと理解いただきたい）であるとされる処遇の結果のなかで，実は成績がふるわなかったひとがいる，という場合も，当然ありうる。図3は，すべて同じ効果量（Δ = 1.00）をもつが，実験群のばらつきが異なる例をあらわす。

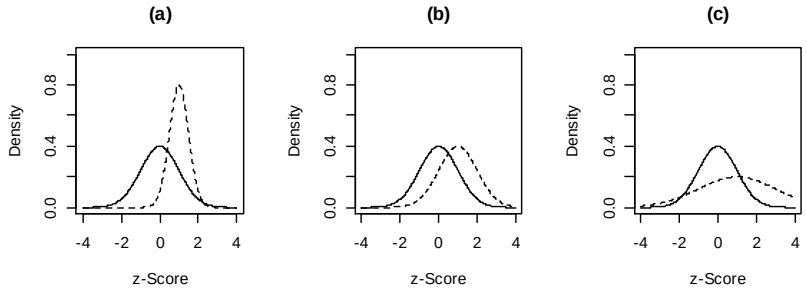


図 3. 同じ標準化平均差（⊿ = 1.00）と異なるばらつきをもつ 3 つの分布の例

（a）の例では，ほぼすべてのひとが，統制群における平均点のレベルよりも高い成績をとっているが，（c）では，30%程度のひとが統制群の平均点のレベルよりも低い成績をとっている。もしも，事前の成績において 2 群に等分散性があったなら，（a）は，処遇によって，ばらつきが小さくなったといえるし，（c）は逆に大きくなったといえる。このように，処遇が成績のばらつきにおよぼす影響は，教育的な観点のうえで重大な意味をもつ（e.g., 前田, 2008; 荻野, 1983）。

基本的には，処遇によって成績のばらつきが大きくなると，平均差が高くても，逆

に大きく成績が下がったという人がいたり，なにかしらの適性処遇交互作用（ATI）があったということが考えられる（e.g., 前田, 2008）。また，平均差が 0，つまり効果量が 0 であっても，データが処遇の結果として特徴的な性質を示す場合も多い。図 4 に例をあげる。

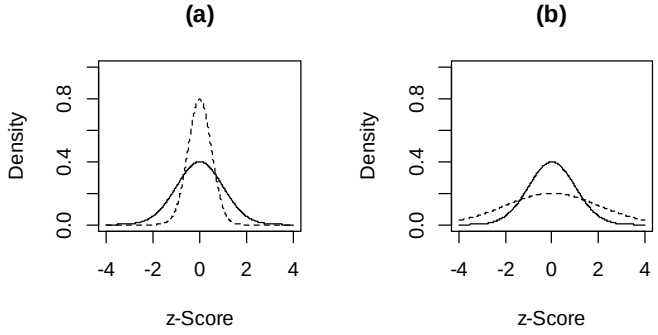


図 4. 平均差がともに 0 であるが異なるばらつきをもつ 2 つの分布の例

図 4 における（a）の例では，成績下位層の点数が向上しているともいえるし，（b）では，逆に低下しているともいえる。上位層では，その逆である。例の（b）の方が，なにかしらの ATI があった可能性が高いため，ATI に関する情報（ほかの要因）を得て，学習者の細分化（learner segmentation）をおこなうとよいかもしれない。このことは，図 3 のように，効果量が 0 でない場合についても同様である。

処遇の結果として生じた，ばらつきの差異といった情報は，もちろん，中心傾向のみをあらわす効果量からは基本的に得られない。よって，効果量のみをもって，処遇の結果を適切に解釈するためには，重要な情報が不足してしまう。

3.6 教育従事者の関心が分布の中心傾向に限らないこと

先の点にも関わるが，これは，著者が想定する実務的な観点にかぎってのことである。教育従事者の意思決定の重心は，分布の中心傾向，特に平均値のみに依拠しないであろうし，するべきでない場面もあると考えられる。

もちろん，個々の傾向を代表する値としての平均値は，もっとも重要である。平均値は，集団の各ケースにおける偏差自乗の平均値（= 分散）を最小化する値であるのだから，個々の成績とのズレが，総合的にみて，もっとも小さくなる値といってよい。しかし，教育実践における重要性はその中心傾向だけに限るわけではない。

教育従事者は，ときに，平均的な成績の向上よりも下位層の成績の向上を重視するであろうし，逆に，下位層よりも，上位層における成績の向上を重視する場合もあるかもしれない。たとえば，なんらかの補修授業などに類するカリキュラムでは，集団の平均値よりも，下位の成績，極論すれば，もっとも芳しくない成績をとったひとの得点が処遇の望ましさをあらわしてもよい。さらに，極端な例になるが，留学希望者に対するプログラ

ムに 100 人がいて，そのなかで，実際に留学する資格を得るのは，上位 5%であるとする。このときもやはり，平均値の向上（効果量）よりも，その上位 5%の能力の伸張に意思決定の重点があったとしても，なんら非合理的ではない。もちろん，そのような場合は，目標規準準拠にもとづくほうが適切であるかもしれない。しかし，この例に，集団基準準拠のような側面がないとはいえない。

さらに，ミクロ経済学があきらかにしているように，消費者の効用は財の追加にともない，逓減していく場合もある。そのため，教育従事者の意思決定の重心が，上位層の伸びよりも下位層の伸びのほうに，やや歪んだものだとしても不思議ではない。また，学習の高原現象（pleatau phenominon）ということもあり，上位層を伸ばすコストよりも，下位層を伸ばすコストのほうが低くなる可能性もある。第二言語習得理論でも，言語運用技能の発達は，労力ないし時系列に対して線形であるとは限らないと考えられている。このような場合，下位層における成績の変化のほうが，上位層のそれよりも，期待価値が相対的に高いという場合もありうる。もちろん，場合によってその逆もありうる。いずれにせよ，教育実践に関わる実務のなかで，実施のコストといった，無数の要因を黙殺して，盲目的に中心傾向のみに注目することは，あまり現実的でないようにおもわれる。くりかえしになるが，だからといって，中心傾向の情報が不必要であるとか，重要ではないということには，けっしてつながらない。

具体的な例を図 5 にあげる。

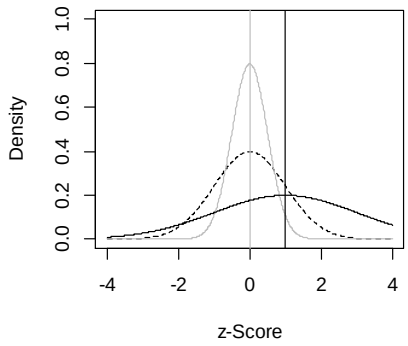


図 5. 効果量と実験群のばらつきが異なる分布の例

図 5 では，破線が統制群の分布，黒と灰の実線がふたつの実験群の分布をあらわしている。黒の分布はΔ = 1.00 であるが，ばらつきが非常に大きい。一方，灰の分布は効果量が 0 であるが，ばらつきが小さい。通常，効果量大といわれる黒の分布よりも，効果量が 0 である灰の分布の方が，あきらかに下位層が少ない。この場合，効果量が大きい黒の分布を生む処遇のほうが望ましい，と一概にいえるであろうか。処遇の結果の望ましさを，一種の効用（utility）であるととらえると，おろらく，統計的な効果は効用に影響するひとつの因子でしかないだろう。

3.7　有益な分布の歪みに関する情報を反映できないこと

そもそも，代表値としての平均値の頑健性，という問題にもつながるが，標準化平均差などの効果量は，データの正規性が満たされない場合，値が妥当でなくなる場合がある。図6には，平均値と標準偏差が等しいふたつの分布の例を示す。実線，破線ともに，$M = 30$, $SD = 10$である。

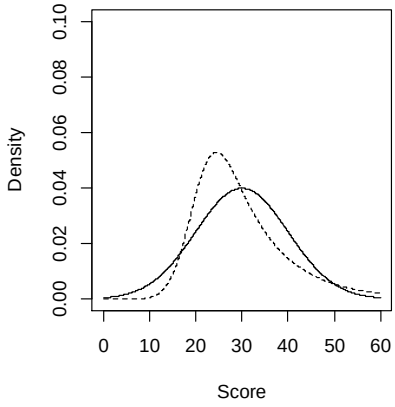


図 6. 平均値と標準偏差が等しいふたつの分布の例

歪度をもつデータから算出された効果量は，統計的に望ましくないばかりでなく，意思決定のありかたをも左右する。破線のデータも $M = 30$ であるが，中央値はこの値でないため，半分以上のひとが平均値にみたない。これでは，半数以上のケースが，効果量であらわされる処遇の結果よりも，芳しくない成績をとったということになる。こうしたデータに対して，効果量は，集団に対する処遇の結果を適切に反映できているといえるのだろうか。また，破線のデータは，実線のデータよりも下位層のひとが少ないが，ふたつのデータの標準偏差は，等値であることにも注意していただきたい。

くわしくみると，破線のデータは，平均よりも下位層と，平均よりも上位層における分布の勾配が異なる。左側の裾のほうが，急で，右側の裾が緩やかである。このことは，下位層に，なにかしら特段の配慮を施し，本来あるべき（対称であるべき）ばらつきを，人為的に狭めた結果を示しているという可能性がある。また，対象とする特性における変化の段階が非線形であることが理由かもしれない。効果量は，こうした特徴的な分布の情報も反映しない。

いずれにせよ，外国語教育に関わる限り，学力，言語知識，言語技能といったものの分布が，かならずしも集団において正規性をもつとは限らない。効果量は，平均値と標準偏差のみからもとまるため，必然的に，分布の歪みについての情報は捨象しなければならなくなる。しかし，実務的な観点のもとで，こうした分布の歪みについての情報が，概して不要だとはいえないだろう。そういった分布の歪みが，処遇の結果の望ましさに影響していてもおかしくない。

3.8 個々のケースについての理解が困難であること

対応がある場合と，対応がない場合では事情が異なるが（後述），効果量のみによる処遇の結果の解釈では，集団の傾向こそ把握できるものの，個々の成績の振る舞いについて理解することは困難である。学術的な観点では，個人は母集団から標本化された1ケースに過ぎないが，教育実践に関わる限り，集団の振る舞いに主な関心があっても，個々の評価が関心から完全に外れるわけではない。

むしろ，望ましくは，集団の傾向と個々の傾向を，関心に応じて，適切なレベルで集約することである。これまでのべてきたように，処遇をうけた集団の傾向としての効果量は，集団がもちうる情報を，以下の点で捨象しすぎている：(a)単純効果量である平均差や⊿などの標準化平均差のみによる解釈は，処遇が生んだばらつきを考慮しない，(b)標準化平均差は分布の中心傾向のみをあらわす，(c)平均差や標準化平均差は，有益な情報でありうる，分布の歪みを反映しない。だからといって，データの要約をせずに，個々の振る舞いにのみ関心を寄せていては，一般化は困難なものになるし，その解釈には，さらに莫大なコストがともなう。

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には，(a)集団の中心傾向のみでなく，個々のケースの振る舞いをある程度反映し，(b)より，さまざまな文脈に対応できるという点において柔軟であり，(c)情報を適切なレベルに集約することができ，(d)そしてなによりも，解釈が容易な定量的方法こそが望まれる。以降では，そうした方策のありかたを議論する。

4. 効果量を解釈するための工夫

この節では，まず，効果量の値の解釈について，その方策の一例を示すために，これまで提案されてきた効果量の解釈のための方策を紹介する。本論は，効果偏差値（伊藤, 1998）と優越率（probability of dominance; 南風原・芝, 1987；南風原, 2014）ないし共通言語効果量（the common language effect size; McGraw & Wong, 1992）を対象とする。

4.1 効果偏差値

効果偏差値という用語は，伊藤（1988）によるものであるが，この考えかた自体は非常に基本的なものである。伊藤は，一般的に，標準化得点（z 得点）のスケールをイメージできるものは少ないという点に注目し，実用的な観点から，これを偏差値（standard score）のスケールに規格化することで，解釈が容易になるのではないか，とのべている。

前述したが，そもそも，標準化得点とは，得点を平均 0，標準偏差 1 となるように変換したものである。この変換は，標準化ないし z 変換ともよばれるのであるが，標準化したデータの分布が，かならずしも，正規分布にしたがうとはかぎらない。

標準化された個々の得点を，標準化得点ないし z 得点とよぶ。n 人のあるグループ 1 に属するケース i の得点 x_i を，標準化した（標準化）得点 z_i は，

$$z_i = \frac{x_i - \mu_1}{\sigma_1}$$

であたえられる。これは，標準化した得点分布において，

$$\frac{\sigma_1}{\sigma_1} = 1$$

であり，さらに，

$$\sum_{i=1}^{n}(x_i - \mu_1) = 0$$

であると考えるとイメージしやすい。

偏差値は，日本では，比較的馴染みやすいものかもしれないが（伊藤，1998）。これは，標準正規分布 N（0, 1）ではなく，N（50, 10）に変換したものである。よって，

$$\text{偏差値}_i = 10z_i + 50$$

と計算できる。

さて，統制群をグループ 1，実験群をグループ 2 とすると，Δは，

$$\Delta = \frac{\mu_2 - \mu_1}{\sigma_1}$$

であった。これは，実験群の平均値を，統制群の平均値と標準偏差で標準化したものととらえられる。この操作を，本論では便宜的に比較標準化とよぶ。そして，比較標準化された統制群の個々のケースの平均はΔに等しい。冗長なようだが，Δはこのようにも書くこともできる。

$$\Delta = \sum_{i=1}^{n_2}\left(\frac{x_{2i} - \mu_1}{\sigma_1}\right)$$

繰り返しになるが，このことから，Δは，「統制群のデータで標準化（比較標準化）した

個々のケースにおける標準化得点の平均値」ともみることができる。これを偏差値のスケールに変換すると，「統制群のデータと比較した場合の，実験群の平均的な偏差値」と解釈できる。また，Δを変換した値は，集団の偏差値をあらわすもの，とも理解できる（伊藤, 1988）。仮に，$\Delta = 0.20$ であれば，統制群と比較して，実験群の成績は平均的に，偏差値 52 程度であったと解釈できる。これは，標準化得点のスケールになじみがなく，そして，偏差値にはなじみがある教育従事者にとっては，有益な方策になりうる。

さらに，標準化得点を確率であらわすこともできる。たとえば，Δが 0.20 であれば，

$$\Phi(0.20) \fallingdotseq .58$$

であるから，実験群の平均値は，統制群の下から 58%程度の位置にあると考えられる。これは，逆にいうと，実験群の平均値よりも，上位の成績を収めた統制群の人は，42%程度だと予測できるということである。ここでの Φ は，標準正規分布における下側確率を返す関数である。参考までに，図 7 に，標準正規分布における累積密度関数を描く。

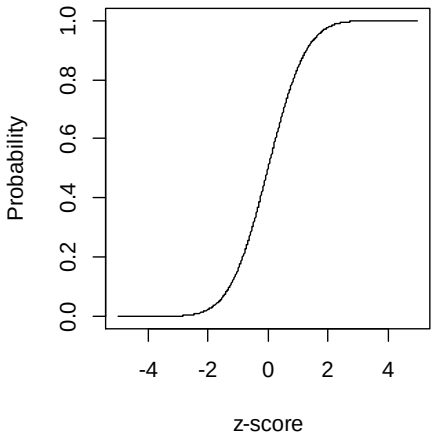


図 7. 標準正規分布における下側累積密度関数

このように，Δなどの標準化平均差を，標準化得点ではなく，偏差値や確率に変換することは，効果量をより解釈しやすいものにするかもしれない。しかし，これはあくまでも解釈のためのものであって，種々の変換自体によって，効果量の値が変わるわけではない。

4.2 優越率と共通言語効果量

優越率（南風原・芝, 1987）と共通言語効果量（McGraw & Wong, 1992）は，基本的に同様のものである（南風原, 2014）。ここでは，刊行時期が，よりはやい優越率という用語をもちいる。

優越率は，南風原・芝が提案した，標準化平均差の解釈指標である。南風原・芝は，行動科学における知見が，そもそも決定的なものであることは少なく，むしろ確率論的な関連の強さについてである場合が多い，という考察のうえで，標準化平均差を，そのまま確率的に解釈することの有効性を主張している。優越率は，標準化平均差に直接対応するものであるが，対応がないデータにおける優越率（π_d）は，「母集団 A から任意に選ばれた被験者の得点 X_A が,それとは独立に母集団 B から任意に選ばれた被験者の得点 X_B よりも大きくなる確率」（p. 70）と定義されている。[2] つまり，

$$\pi_d = \Pr(X_A \leq X_B)$$

である。パラメトリックな方法で，この確率をもとめるには，

$$\pi_d = \Phi\left(\frac{d}{\sqrt{2}}\right)$$

とする（p. 71）。ここでの d は，Cohen's d である。優越率は，それぞれの群から 1 ケースずつ抽出したときに，任意の群の値が高い確率であるともいえる。これは，標準化得点のスケールよりも，直感的な理解を助ける。図 8 に優越率と標準化平均差 d の関係を描く。

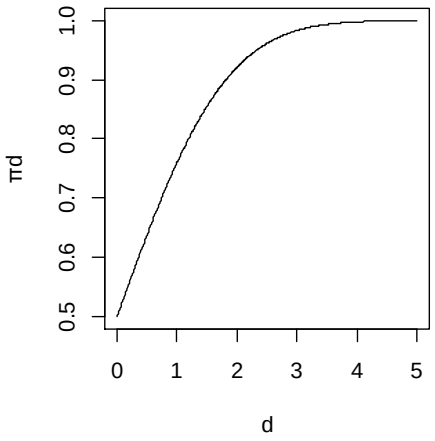


図 8. 対応がない場合の優越率と標準化平均差の関係

事前―事後間の比較など，対応があるデータの場合は，「母集団において X_A のほうが X_B よりも大きい被験者の割合」（p. 72）と考えられる。南風原・芝は，この優越率を π'_d としている。この算出のためには，まず，差得点にもとづく標準化平均差を導入しなければならない。差得点にもとづく標準化平均差は，まず，ケース i の二点間（A，B）または二変数間の差を計算し，

$$D_i = A_i - B_i$$

つぎに，平均差の平均値 μ_D と，差得点の標準偏差を σ_D をもとめる。差得点にもとづく標準化平均差は，

$$d_D = \frac{\mu_D}{\sigma_D}$$

となる。これをもちいて，対応がある場合の優越率 π'_d は，

$$\pi'_d = \Phi(d_D)$$

であたえられる。参考までに，図 9 に，対応がある場合の優越率と差得点にもとづく標準化平均差の関係を示す。

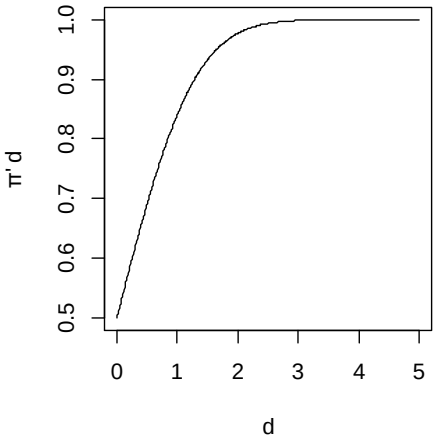


図 9. 対応がある場合の優越率と差得点にもとづく標準化平均差の関係

このような計算を，ノンパラメトリックな手法で代用するならば，対応がない場合は，2 群から標本 1 ケースずつ抽出する際のすべての組み合わせについて計算し，直接計算すればよい。対応がある場合は，実験参加者の総数を分母として，差得点が正の符号をもつケースを数えあげればよい（p. 76）。そのような場合があるかはわからないが，すべての組み合わせについて計算することが，莫大な計算量になる場合は，1 ケースずつの再標本化（ブートストラップ）を任意の数だけ繰り返すことによっても，簡易的な近似値が得られると考えられる。

5. 集団に対する処遇の結果を適切に解釈するために

ここからは，効果量のみを利用して処遇の結果を解釈することの限定性を補償するであろう，いくつかの定量的方法について試案をのべる。本稿が対象とするところは，NEGD における対応がない比較（統制群―実験群）と，対応がある比較（事前―事後）である。まずは，対応がない場合を前提として進めていく。

5.1 ばらつきの比率

第一に，これまでみてきたように，⊿などの標準化平均差は，中心傾向こそ示すものの，実験群のばらつきについての情報はもたない。そこで，比較標準化されたスケールにおける実験群のばらつきを数量的に把握する必要がある。

比較標準化する前の実験群のばらつきは，標準偏差の定義式から，

$$\sigma_2 = \frac{1}{n_2}\sqrt{\sum_{i=1}^{n_2}\left(X_{2_i} - \mu_2\right)^2}$$

である。[3] また，比較標準化したケース i の成績 X は，

$$z_{2_i} = \frac{X_{2_i} - \mu_1}{\sigma_1}$$

である。さらに，比較標準化した実験群のばらつきは，比較標準化した実験群の平均値が⊿であるため，

$$\text{Standardized } \sigma_2 = \frac{1}{n_2}\sqrt{\sum_{i=1}^{n_2}\left(z_{2_i} - \Delta\right)^2}$$

とあらわせる。これを解くと，

$$= \frac{1}{n_2}\sqrt{\sum_{i=1}^{n_2}\left(\frac{\left(X_{2_i} - \mu_1\right)}{\sigma_1} - \frac{(\mu_2 - \mu_1)}{\sigma_1}\right)^2}$$

$$= \frac{1}{n_2}\sqrt{\sum_{i=1}^{n_2}\left(\frac{\left(X_{2_i}-\mu_1\right)-\left(\mu_2-\mu_1\right)}{\sigma_1}\right)^2}$$

$$= \frac{1}{n_2}\sqrt{\sum_{i=1}^{n_2}\left(\frac{X_{2_i}-\mu_2}{\sigma_1}\right)^2}$$

$$= \frac{1}{n_2}\sqrt{\sum_{i=1}^{n_2}\left(X_{2_i}-\mu_2\right)^2}\frac{1}{\sqrt{\sigma_1^2}}$$

となる。ここで,

$$\sigma_2 = \frac{1}{n_2}\sqrt{\sum_{i=1}^{n_2}\left(X_{2_i}-\mu_2\right)^2}$$

であり，実質的に σ は負の値をとらないので，比較標準化された実験群のばらつきは,

$$\text{Standardized } \sigma_2 = \frac{\sigma_2}{\sigma_1}$$

これは，分散比（F）の平方根に等しい。

$$\text{Standardized } \sigma_2 = \frac{\sigma_2}{\sigma_1} = \sqrt{F}$$

よって，比較標準化された実験群の分布は，$N\left(\Delta,\sqrt{F}\right)$であたえられよう。

事前のテストにおいて，実験群と統制群の間に等分散性があると仮定できる場合に，$\sqrt{F}$ が 1 を超えると，統制群にくらべ，実験群は，処遇の結果によって，ばらつきが大きくなったと考えられる。また，逆に 0 を下回るとばらつきが小さくなったと考えられる。効果量（Δ）が比較標準化された集団の中心傾向をあらわすとすれば，$\sqrt{F}$ はその散布度をあらわすといえる。集団に対する処遇の結果を解釈する際に，$\sqrt{F}$ の値を算出することは，通常の記述統計において，データのばらつきを検討することと，まったく同じことである。もちろん，このばらつきの比率について，統計的仮説検定をおこなうとすれば，等分散性のための F 検定をもちいるとよい。

5.2 比較点cの提案

比較標準化された実験群の分布 $N(\Delta, \sqrt{F})$ の情報をもちいると，効果量のみによる結果の解釈よりも，さまざまな場面に応じた，柔軟な分析ができるようになる。たとえば，前述のとおり，処遇のありかたに関連するさまざまな意思決定の場面において，当事者の関心は中心傾向のみに限らない。むしろ，意思決定の場面場面に依存して，相対的に重要視される部分が，分布上で異なりうる。ここでは例として，中心ではなく，むしろ分布の端に注目してみる。

たとえば，比較標準化された実験群の成績における予測区間の下限値は，「処遇を受けた人が取りうる最低限の成績をあらわす標準化得点」をあらわすと解釈できる。このように，比較標準化された分布の情報をもちいて，処遇の結果を検証するために参照する値を，ここでは，総称として「比較点（c）」とよぶ。

例として，95%予測区間の下限を比較点 c とし（下限比較点とよぶ），この値をもとめるには，

$$c = \Delta - t_{97.5}\sqrt{F\left(1 - \frac{1}{n}\right)}$$

ただし，t_α は自由度 n - 1 の t 分布における α 点である。α は，任意の予測区間の片側の点を考えればよい。もちろん，場合に応じて，予測区間の幅を変えて比較点を定めてもよい。予測区間を 95%とするのであれば，簡易的に以下のような式に置き換えても，実務レベルでは支障ない程度の近似値が得られると考えられる。

$$c = \Delta \pm 2\sqrt{F}$$

これは，標準化平均差（Δ）を，実験群のばらつきで重みづけした指標と考えてよい。つまり，効果量がより高く，なおかつ，よりばらつきが狭くなる処遇が c を最大化することになる。図 10 に例をあげる。縦線が下限比較点である。例（a）～（c）の効果量は同一であるが，ばらつきの狭まりにともない，比較点の値は上昇していく。同様に，（d）～（f）は，ばらつきが同等であるが，効果量の増大につれ，比較点の値も上昇していく。また，標本サイズが大きくなればなるほど，推定精度の向上によって値は上昇する。

もちろん，状況によって，予測区間の上限をもとめることが有効である場合もある。ここでは，これを上限比較点とよぶ。上限比較点も効果量の増大にともなって，その値が上昇していき，実験群のばらつきが広がるほど，値が上昇していく。

より頑健な方策としては，予測区間ではなくて，実験群の分布における任意の分位点をもとめてもよいかもしれない。四分位点や，10%点，90%点などが必要になる場合も

ありうる。さらに，比較標準化された実験群のデータにおける中央値や最頻値は，分布が歪んでいる場合，効果量⊿によるものよりも，適切な判断を可能にする場面があるかもしれない。しかし，中央値や最頻値は，完全なデータがなければもとめることができない。

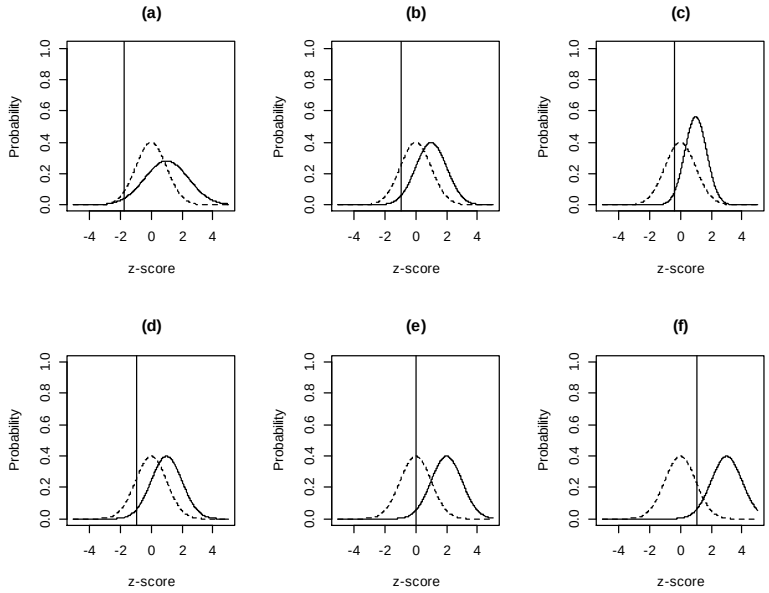


図 10. 効果量とばらつきの変動が比較点の値におよぼす影響

任意の比較点は，標準化得点のスケール上であらわされる。これまでの，標準化得点上における比較点を，ここでは便宜的に c_z と記す。c_z は，標準化得点のスケールになじみのないものにとって，値自体の解釈がむずかしいために，効果偏差値（伊藤, 1988）と同様に，偏差値のスケールに規格化してもよい。偏差値のスケールであらわされた比較点を，便宜的に比較偏差値とよび，c_s と表記する。c_s は，効果偏差値と同様に，

$$c_s = 10c_z + 50$$

でもとめられる。さらに，これを標準正規分布の累積密度確率であらわすこともできる。標準化正規分布は，ここでは，標準化された統制群の分布に等しい。このことから，標準正規分布の累積密度確率であらわされた比較点を，統制群の成績の位置として把握することができる。これを比較分位点（c_p）とよぶ。この c_p は，上側確率の場合，

$$c_p = 100\{1 - \Phi(c_z)\}$$

となる。仮に 95%の予測区間をもちいた下限比較分位点が 50 の場合，実験群の予測区間の下限が統制群の中央値であるのだから，「ある処遇を受けたひとのほとんどの成績は，異なる処遇を

受けたひとの半分よりも上であろう」というように，直感的な見込みをもてる。

また，一般的な解釈基準では，小や大とよばれ，実質的な相対関係の大小が消失してしまうような値の効果量をもつ処遇であっても，比較点をもちいると，適切な意思決定ができる可能性がある。たとえば，実験群および統制群に等分散が仮定できるとして，$\Delta = 3.00$ と $\Delta = 4.00$ の 95%予測区間による下限比較分位点は，$n = \infty$だとした場合，それぞれ 85（%）と 98（%）程度である。これは，0.80 以上の値を，すべて，大としてひとくくりにする解釈基準よりも，場合によっては，適切に，そして，ことこまかに現実を反映するかもしれない。しかし，あきらかに統制群と実験群の分布の重なりがない場合，比較分位点では適切な評価ができない。そのような場合は，分布のばらつきを実質的に捨象できる可能性が高いため，効果量のみによる議論が有効かもしれない。

考えかたがやや異なるものの，標本の予測区間のみではなくて，標本値の信頼区間について似たような手順をもちいてもよい。標本効果量には誤差があるため，標本効果量を母効果量のように解釈してはならない。点推定値でなく，信頼区間の下限を積極的に解釈することは，より保守的で有効な方法であるといえる。

比較点の算出には，文脈に応じてさまざまな方式（e.g., 任意の確率における予測区間，任意の分位点）が考えられるが，その区間推定をおこなう必要性もあろう。そのようなときは，簡便なブートストラップ法をもちいてもよい（草薙, 2014b）。

このように，本稿が提案する比較点は，中心傾向のみを示すΔよりも，場合に応じて，豊富な情報をもたらす可能性がある。また，中心傾向に縛られず，実務的な観点のもと，さまざまな文脈や状況に応じて，自由に任意の点に着目することは，この手法に限らず重要なことである。本稿の付録に，比較点の計算例をあげている。

5.3 対応があるデータにおける基本

これまでは，基本的に対応がないデータを前提として論を進めてきた。しかし，NEGD における事前―事後の比較，または NEVD における比較は，対応があるデータの比較となる。対応があるデータの比較は，そのデータの対応関係，つまり共分散を考慮する必要があるので，対応がないデータの比較とは，一部手法が異なる。

もちろん，対応があるデータを，対応がないデータのようにあつかうこともできる。しかし，これはさまざまな面で情報の損失がある。図 11 は平均値と標準偏差が等しいが，共分散が異なる分布の例である。つまり，これらの例は，共分散が異なるものの，Δなどの一部の標準化平均差の値はまったく同一である。

(a)の例では，およそ $r = 0$ である。これは，処遇の結果としては，考えにくいことであるが，事後の成績が事前の成績から予測できない状態である。また，なにかしら，成績に関連する構成概念とは異なるものに強い ATI があるのかもしれない。さらに，個々のケースにおける伸び幅（差得点）のばらつきが大きいともいえる。一方，(c)の例は，逆の傾向を示している。伸び幅のばらつき

は小さく，おそらく，ATI に関連する現象と関わりが深そうなデータではない。例の(b)は，いずれもその中間の状態を示している。

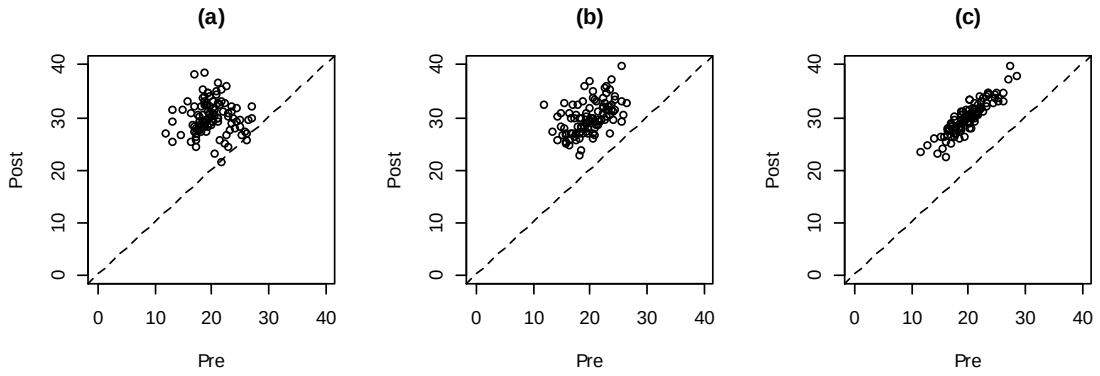


図 11. 平均値と標準偏差が同一であるが，共分散が異なる分布の例

伸び幅のばらつきについて，もう少し注目してみる。まず，図 12 に，図 11 と同じデータについて，今度は伸び幅のヒストグラムを描いてみる。図 12 から一目瞭然であるが，やはり，共分散の違いによって，伸び幅のばらつきがあきらかに異なることがわかる。

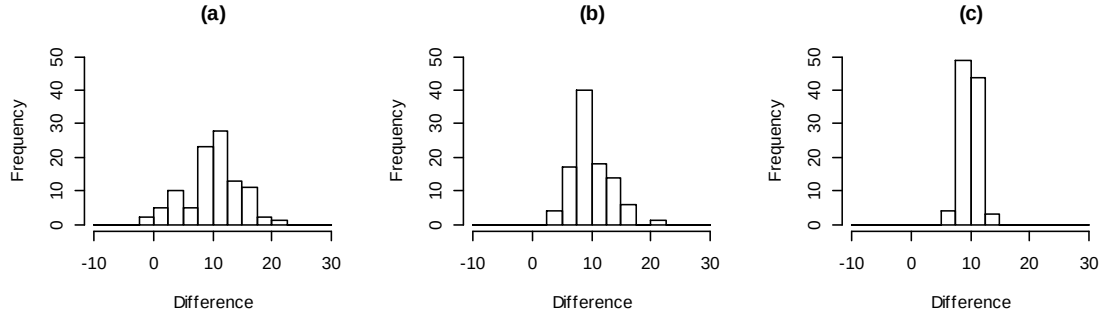


図 12. 平均値と標準偏差が同一であるが，共分散が異なるデータの差得点のヒストグラム

実は，これは，共分散が 2 変数の偏差積の平均値であることから自明である。厳密ではないが，これは，以下のように考えるとわかりやすい。個々のケースにおける偏差積は，両変数の偏差が同符号で，さらにそれぞれの変数の偏差が大きくなればなるほど，大きい値をもつ。共分散は偏差積の平均なのだから，異符号の偏差をもつケースが多ければ，共分散は小さくなる傾向にあり，逆に同符号の偏差をもつケースが多ければ，共分散は大きくなる傾向にある。2 変数の値の差が小さければ小さいほど，同符号の偏差をもつ確率が高まるのであるから，共分散が大きければ大きいほど，差のばらつきは小さくなる。

伸び幅のばらつきは，教育実践のなかで，効果量と同じように大きな意味をもちうるし，伸び幅のばらつきを考慮する必要があるときは，先に紹介した，差得点にもとづく標準化平均差をもち

いるとよい。また，差得点のヒストグラムを確認することも肝要である。いずれにせよ，対応があるデータのときは共分散についても考慮し，論文などでは，積極的に報告すべきである。

先に対応がないデータのついての論でのべた，ばらつきの比率（F やその平方根）を検討することは，対応があるデータの際も有効である。仮に共分散が同一であったとしても（仮に 0 としても），ばらつきの比率が異なれば，全体としてのデータの傾向が大きく変わる。これは図 13 の例からもわかる。

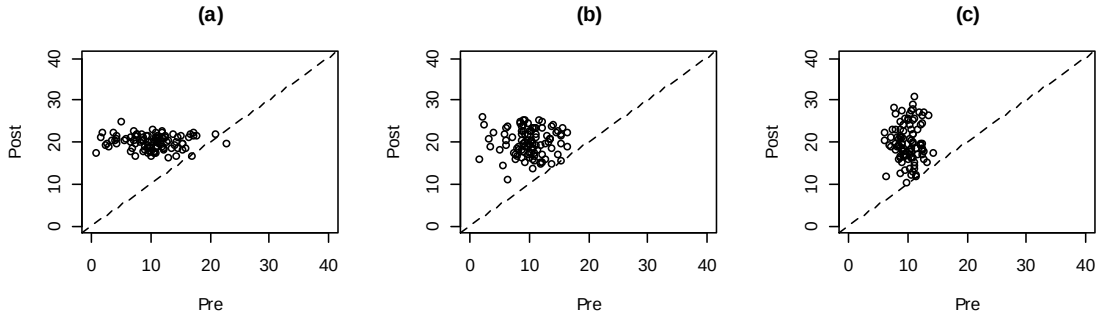


図 13. 平均値と共分散が同一であるが，ばらつきの比率が異なるデータの散布図

3 つの例における共分散はすべて 0 である。図 13 の（a）は事前テストのばらつきが非常に大きいが，事後でばらつきが小さくなっている。（c）はその逆で，（b）は事前―事後で同等である。このような特徴を見逃さないためにも，対応があるデータの分析においても，ばらつきの比率を検討することが重要となろう。図 13 の例では，簡略化のために，共分散をすべて 0 としたが，処遇の結果を比較する際に，共分散が 0 であるということは想定しにくい。共分散とばらつきの比率には複雑な連関があるため，実際のデータでは両方を考慮する必要がある。

5.4 対応があるデータにおける比較点

対応があるデータにおいても，対応がないデータと同様に比較点をもちいることができる。しかし，対応がないデータと同様の処理をすると，先にのべたように，重大な情報の損失につながってしまう。まず，もっとも簡便な方法としては，単純効果量である平均差の信頼区間をもとめて，その下限や上限を検討するとよい。また，差得点の予測区間を検討することも，きわめて重要である。

場合に応じて，「母平均差が 0 である」という帰無仮説に対する対立仮説に関心がある場合は，t 検定をしてもよい。これは差得点にもとづく母効果量が 0 であるかについて検討することと同じである。このほかの方法としては，以下に示す分位点回帰をもちいると，適切かもしれない。

5.5 分位点回帰をもちいた分析

一般的に，対応があるデータにおいて，中心傾向のみに縛られない分析をするためには，分位点回帰（quantile regression）をおこなうとよい。分位点回帰は，線形回帰モデルのひとつであるが，従属変数の平均値を予測する従来の回帰分析と異なり，任意の分位点（およびパーセンタイル）を予測対象とすることができる（e.g., Hao & Naiman, 2007; Koenker, 2005）。国内でも，分位点回帰は経済学，社会心理学などで，その手法的有効性が指摘されている（e.g., 石黒, 2013）。また，言語テスト研究でも，その応用可能性が主張されている（Chen & Chalhoub-Deville, 2014）。

分位点回帰は，従来の中心傾向のみに依存する分析とは異なり，さまざまな統計的制約から自由であり，分布の歪みや外れ値の影響に対しても頑健である。また，複数の独立変数における影響の強さを，さまざまな分位点間で比較することもできる（同時分析）。ここでは，集団に対する処遇の結果を，対応があるデータから解釈する場面において，分位点回帰が，どのように有効な方策となりえるかを考察する。

まず，対応がないデータと同様に，事前のデータの平均と標準偏差をもって，事後のデータを比較標準化する。比較標準化された事後のデータは，$N(\Delta, \sqrt{F})$の分布に，標準化した事前のデータは，標準正規分布にしたがう。このデータセットを対象として分析をする。

表 1 にあるデータを例としてとりあげる。これは，*R*（R Core Team, 2014）と，パッケージ *MASS*（Venables & Ripley, 2002）を使用して，多変量正規分布にしたがう乱数を作成したものである。

表 1

分位点回帰をもちいた分析で使用するデータの記述統計（$N = 100$）

	平均	標準偏差
事前	20.00	5.00
事後	30.00	5.00

なお，両方のデータが正規分布にしたがっており，共分散は 12.50 であり，$r = .50$ である。ちなみに，このデータの効果量は，単純効果量である平均差が 10，標準化効果量は，$\Delta = 2.00$ である。差得点にもとづく標準化平均差も，2.00 である。これらを効果偏差値であらわすと，70 となる。また，優越率 π'_d は，.98 である。

最初に，F が 1 であることから，処遇によって，ばらつきに変動はなかったと考えられる。つぎに，対応がない場合と同様に，95%予測区間にもとづく下限比較点 c_z をもとめる。c_z は，およそ 0.07 である。

これを，比較偏差値であらわすと，$c_s = 51$，比較分位点であらわすと，$c_p = 53$ である。このことから，ほぼ全員の事後の成績は，事前の成績の平均を上回ると推測できる。図 14 に標準化した事前の成績，比較標準化した事後の成績の散布図を描く。なお，斜めの破線は $y = x$，横に走る実線は標準化平均差，点線は下限比較点をあらわしている。

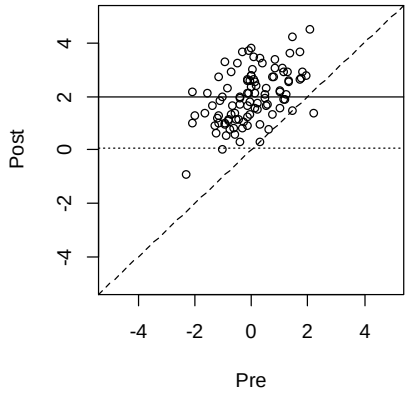


図 14. 標準化平均差および下限比較点を付記したデータの例

図 14 をみると, 優越率や比較点がデータとほぼ完璧な整合性を示していることがわかる。ここまでは，対応がないデータの分析と同様である。ここからは，分位点回帰をもちいる。まず，下側 5%点を予測の対象として（τ = .05），分位点回帰をおこなう。R のパッケージ *quantreg*（Koenker, 2015）をもちいた分位点回帰の結果,

事後の成績 = 0.49×事前の成績 + 0.50

という回帰式が得られた。切片と係数の 95%信頼区間は，表 2 にまとめている。図 14 に対して，この回帰式を灰色の実線で描き入れると図 15 のようになる。

表 2

分位点回帰（τ = .05）の点推定値および 95%信頼区間の下限と上限

	点推定値	95%信頼区間の下限	95%信頼区間の上限
切片	0.50	0.28	0.83
係数	0.49	0.24	0.64

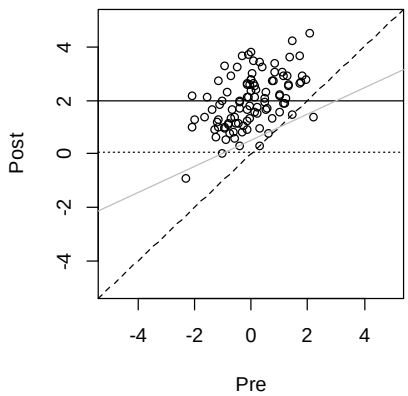


図 15. 標準化平均差，下限比較点および分位点回帰直線を付記したデータの例

かならずしも厳密ではないが，実務的には，ここで得られた回帰式に，対応があるデータにおける下限比較点と同じような解釈をあたえてもよい。つまり，処遇を受けた人が取りうるだろう成績の下限ラインである。もちろん，対象とする分位点は任意であり，下側 5%に限らない。教育実践の文脈では，さまざまな分位点が興味の対象となりうる。当然ではあるが，95%の予測区間の下限値も，下側 5%の分位点も，分布の端側をあらわすにすぎないのであり，統計的に同一のものではないし，5%というような値を固定的なものとしてとらえるべきでない。

もちろん，得られた回帰式から，事前の成績の位置に対応する任意の分位点における予測値をみちびくことができる。このは，効果量よりも，はるかに自由度の高い解釈を可能にするかもしれない。たとえば，複数の分位点を対象とする同時分析から，表 3 のようなデータが得られる。ここでは，上記の例を題材にしている。切片は，事前の成績が平均点程度であるひとたちの，事後における任意の分位点を示すと考えられる。たとえば，τ = .50 は中央値であるのだから，このときの切片 1.94 は，通常の効果量に対する，やや頑健な推定値としてみることができよう。また，各分位点における係数を比較することが重要である可能性もある。

表 3

5%点，25%点，50%点，75%点，95%点を対象とする分位点回帰の例

		点推定値	95%信頼区間の下限	95%信頼区間の上限
τ = .95	切片	3.49	3.37	3.78
	係数	0.51	0.09	0.79
τ = .75	切片	2.60	2.36	2.77
	係数	0.58	0.22	0.71
τ = .50	切片	1.94	1.71	2.12
	係数	0.43	0.36	0.73
τ = .25	切片	1.50	1.30	1.56
	係数	0.50	0.29	0.73
τ = .05	切片	0.50	0.28	0.83
	係数	0.49	0.24	0.64

ここで，例をあげきることはできないが（付録を参照のこと），分位点回帰をもちいた分析は，データがもつ共分散や，ばらつきの比率，分布の歪みなど，これまで述べてきた「効果量のみによる処遇の結果の解釈」に内在する問題の大部分をカバーしており，さらに中心傾向に束縛されない自由な分析を可能にする。この点において，外国語教育における実務的な観点との親和性が，十分に高いと考えられる。もちろん，ここでは，標準化得点のスケールであらわされているが，値を偏差値のスケールなどに変換して解釈することも重要である。

5.6 目標規準準拠の考えかたと比較点および分位点回帰

個々の成績評価と集団に対する処遇の効果測定は，けっしておなじものではない。しかし，指導法や教材の選択は，個々の成績評価ではなく，集団に対する処遇の結果の解釈に依拠することが多い。このとき，効果量は集団の，まさに中心傾向のみに依拠する方法であり，解釈も容易であるとはいえない。しかし，本稿が提案する，比較点や分位点回帰をもちいた分析の方策は，効果量のみを利用する，集団に対する処遇の結果の解釈よりも，いくつかの特定の文脈において優れているといえるところがある。

通常，目標規準準拠の考えかたでは，集団基準準拠の場合よりも，相対的に個々の振る舞いに関心がある。このようなとき，中心傾向のみをあらわす効果量よりも，分布上の任意の点を選択できる比較点や分位点回帰をもちいた方法は，上位層や下位層の伸びを，より適切に定量化できる。そうして定量化された情報は，集団の中心傾向のみに依拠する効果量よりも，個々の振る舞いについて，よりくわしい情報をもたらすと考えられる。

特に，対応があるデータの場合であるが，処遇の結果として生じる成績の分布が，正規分布に従うと限らないのは，目標規準準拠の考えかたのほうである（e.g., 梶田, 2006）。正規分布を逸脱するデータにおける中心傾向（平均値）は，適切な代表値とはいえない。このような場合，分位点回帰をもちいた分析のほうが，効果量のみをもちいて，集団に対する処遇の結果を解釈する方法も，より頑健であるといえる。

6. 総括

本稿では，これまで，集団に対する処遇の結果を解釈する際の効果量が，おもに（a）外国語教育の実務的な観点において解釈が困難なこと，（b）中心傾向のみをあらわすこと，に由来するさまざまな問題点を抱えることを指摘し，その解決策となるいくつかの定量的方法について紹介した。まず，（a）に関する点については，効果偏差値および優越率など，標準化平均差の解釈指標を導入することの有用性をのべた。また，（b）については，比較点，および分位点回帰をもちいた分析法を提案した。さらに，外国語教育の教育実践のなかで，実務家の関心が中心傾向にかぎらないことを，くりかえし強調した。

もちろん，中心傾向にとらわれない処遇の結果の解釈をすすめることは，中心傾向としての平均値や標準化平均差を吟味することが重要でない，ということをまったく意味しない。また，中心傾向に注目した従来の研究報告やメタ分析によって，これまで築きあげられた知見を疑うものにはなりようがない。学問の発展を願わないものは少ないだろうし，みずからの教育実践を他者と共有することは，教育研究の本質であり，分析において，中心傾向を吟味することがその主翼を担うことは当然である。

しかし，効果量を分析に導入することは，外国語教育に，なにをもたらしたのだろうか。外国語教育の関係者から，「有意差には意味がなくて，効果量があればよい」，「効果量が大きいから，指導の効果が大きかった」，「効果量が小さいから，私のデータはダメ」

といった声が聞こえてくるときがある。繰り返し書いているように，統計的有意性が得られないときの効果量報告が意味するのは，大概の場合，単純に，現在の標本サイズが適切でない，ということであり，統計的な意味での効果は，処遇の効果と同じではなく，効果の大きさと知見の重要性は普通，独立しているととらえるべきである。

外国語教育において，効果量は現在，なぜか研究の公表という文脈と，セットとしてとらえられることが多いようである。確かに，学術的な観点において，効果量の重要性は疑いようがないし，学会などがそれをすすめているという面もある。しかし，外国語教育，そして，その研究のありかたは，本質的に効果量から得られる知見のみを対象とするわけではない。

もしかしたら，大きい効果量を観測したとして，嬉々として報告される教育実践の，その対象となる教室のなかで，処遇との相性で，成績が伸びなかった何人かがいるかもしれない。もしかしたら，その逆に，効果量がなかったとして，共有される場が得られなかった，ある教育実践のなかにも，大きく点数を伸ばした数人のひとがいるかもしれない。教育に関する研究では，そのようなひとが大きな関心になってもよい。そのような処遇のありかたが，議論の的になってもよい。しかし，効果量を重視すること，というよりも，効果量のみに依存すること，それは，中心傾向のみに依存することでもあるが，それが，かえってこうした教育従事者の普通の考えを妨げるのであれば，おそらく，慎重に考えなおさなければならないことであろう。

残念ながら，本稿の目的は，そうした状況を解決することですらない。それに，著者が提案するいくつかの方法を利用することを，外国語教育の関係者に強くすすめるものでもない。比較点や分位点回帰をもちいた分析は，単なる一例にすぎないのである。しかし，こうした解釈の試みや，中心傾向のみにとらわれない分析は，効果量というものが，外国語教育になにをもたらしたか，効果量によって，わたしたちの分析のありかたはどうなったのか，といったことを考えるひとつの材料になると考えられる。

定量的方法をもちいることは，手法上の制約にしたがって，人間の自然な，そして現実的な意思決定のありかたから遠ざかることではないし，情報を過剰に集約して，中心傾向のみに注目することでもない。それはおそらく，定量的方法の本質ではない。定量的方法のなかには，効果量ではみれないものをみる方法もある。効果量が重要だからこそ，効果量だけではみえないものにも気を配りたい。本稿には，そうした考えの足がかりになってほしい。

注

1. たとえば，各群のプールされた標準偏差をもちいる標準化平均差（Cohen's *d* など）は，Δの場合のように実験群の散布度を保持しないため，本稿が提案する比較点などを直接的に同じ方法でもとめることはできない。しかし，プールされた標準偏差は，実験群のばらつきが統制群のものよりも大きいとき，その値は低下し，逆に小さいときに

上昇する。これは，⊿とは異なる性質であり，NEGD の場合でも *d* をもちいる利点となりえる。

2. 優越率は，等分散性の仮定を要求する。しかし，野口（1989）は，ここでどのように等分散性を保証するかが不明であると指摘している。
3. ここでは，便宜的に標準偏差を母集団の標準偏差として計算している。不偏標準偏差推定値をもちいることのほうが望ましいが，ここでの計算上は差異がない。

参考文献

Chen, F., & Chalhoub-Deville, M. (2014). Principles of quantile regression and an application. *Language Testing*, *31*, 63–87.

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

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

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 (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.

南風原朝和・芝祐順 (1987).「相関係数および平均値差の解釈のための確率的な指標」『教育心理学研究』*35*, 259–265.

Hao, L. & Naiman, D. Q. (2007). *Quantile regression*. Newbury Park, CA: Sage Publications.

石黒格 (2013).「社会心理学データに対する分位点回帰分析の適用 : ネットワーク・サイズを例として」『社会心理学研究』*29*, 11–20.

伊藤武彦 (1998).「実践的・探索的研究の効果測定における「効果偏差値」の提案」『和光大学人間関係学部紀要』*3*, 15–23.

梶田叡一 (2006).「教育評価入門―学びと育ちの確かめのために―」共同出版.

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

Koenker, R. (2005). *Quantile regression*. Cambridge: Cambridge University Press.

Koenker, R. (2015). quantreg: Quantile Regression. R package, version 5.11. http://CRAN.R-project.org/package=quantreg

草薙邦広 (2014a).「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 (2014b).「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.

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). Examining the difference that robust statistics can make to studies in language acquisition. *Applied Linguistics*, *31*, 368–390.

前田啓朗 (2008).「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.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*,361–365.

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

野口裕之 (1989).「教育心理学に於ける測定・評価研究の動向」『教育心理学年報』*28*, 115–124.

荻野忠則 (1983).『教育評価のための統計法―心と技術の再統合―』日本文化科学社.

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

R Core Team (2014). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna: Austria. http://www.R-project.org/.

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.

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S. 4th Edition*. Springer, New York.

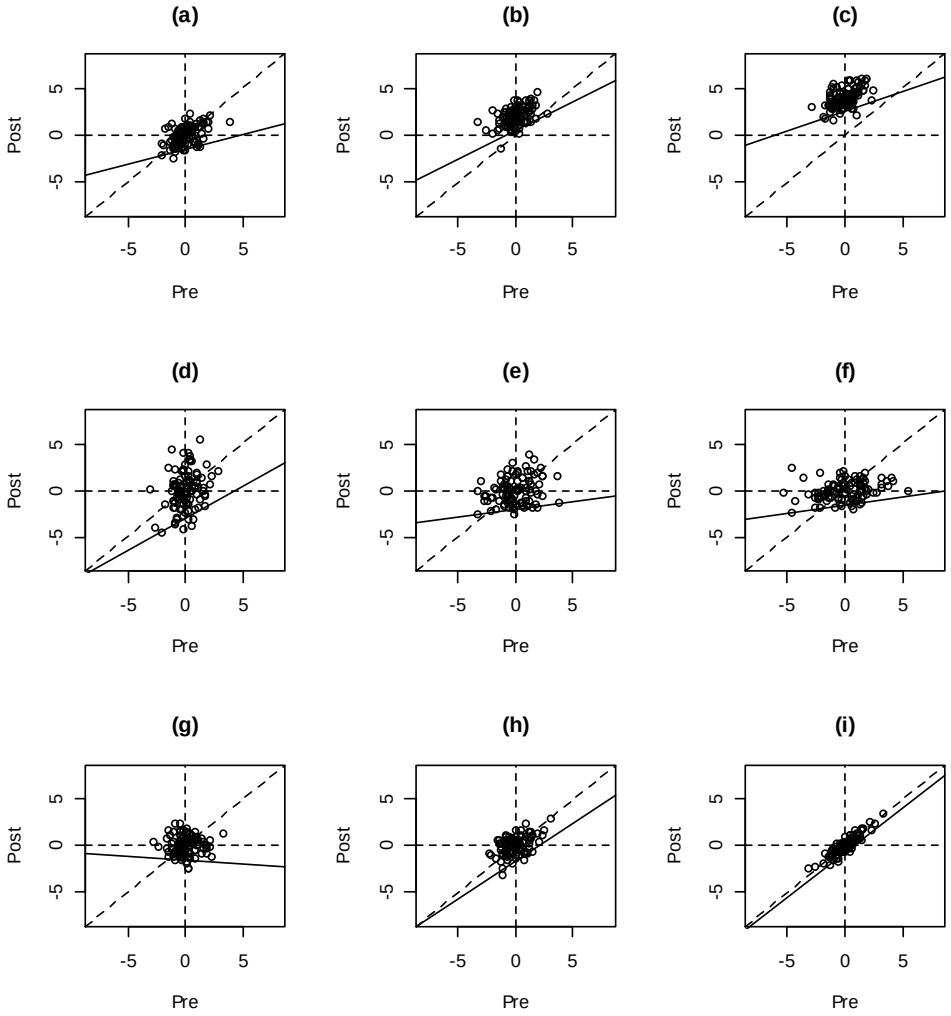
亘理陽一 (2014).「メタ分析の手続きについて：大切なことは全て効果量が教えてくれるか？」『外国語教育メディア学会関西支部メソドロジー研究部会報告論集』*5*, 64–74.

付録

付録１：Ｒにおける下限比較点の計算例

```
cpoint<-function(nc,ne,mc,me,sdc,sde,pi=.975,plot=T){
       x<-seq(-8,8,.01)
       mdif<-me-mc
       delta<-mdif/sdc
       f<-sde/sdc
       cpz<-delta-(qt(pi,ne-1)*f*sqrt(1+(1/ne)))
       cp<-1-pnorm(cpz)
       cprank<-floor(cp*nc)
       if(cprank==0){
              cprank<-1}
              else{
              }
       if(plot==T){
              plot(x,dnorm(x),xlab="Standardized
       Score",ylab="p",type="l",ylim=c(0,2))
       lines(x,dnorm(x,delta,f),col=4)
              abline(v=c(delta,cpz),col=4,lty=2)
       }else{
       }
       list("Mean
Difference"=round(mdif,2),"Delta"=round(delta,2),"F"=round(f,2),"Comparison
score"=round(c(cpz,cpz*10+50),3), "Comparison
Percentile"=round(cp,2)*100,"Comparison Rank"=cprank)
}
```

付録 2：Δ，$\sqrt{F}$，および共分散が $\tau = .05$ の分位点回帰直線にあたえる影響



パラミター設定表

	Δ	$\sqrt{F}$	r
(a)	0.00	1.00	.50
(b)	2.00	1.00	.50
(c)	4.00	1.00	.50
(d)	0.00	4.00	.10
(e)	0.00	1.00	.10
(f)	0.00	0.25	.10
(g)	0.00	1.00	.00
(h)	0.00	1.00	.50
(i)	0.00	1.00	. 90