反応時間分析における外れ値の処理

大久保街亜[1]

Outliers in reaction time data : Methodological considerations and practical suggestions

Matia Okubo

Abstract : Reaction time is used to measure various types of human performance such as perception, memory, and problem solving. Many constructs, from unconscious prejudice to intelligence to personality, can also be measured by use of reaction time. It is, therefore, fundamentally important to remove the influence of spurious long reaction time in a positively skewed distribution, which reaction time data tends to follow. The present article evaluated methods for dealing with reaction time outliers. These methods were categorized into three types : Sample selection, transformation, and whole distribution analysis. In this article, I summarized pros and cons of these methods and made suggestions for a practical reaction time analysis.

Key words : reaction time analysis, outliers, psychometrics

心理学研究における反応時間

反応時間は，さまざまな心理学の研究で使用される反応指標である。古くは18世紀に Wilhelm M. Wundt や Oswald Külpe らも使用した心理学では伝統的なものだ。ただし，現在の心理学研究おける流布は，1960年代に行われた Saul Sternberg による一連の研究に負うところが大きい。Sternberg は，オランダの眼科医学者 Franciscus C. Donders が開発した減算法（Donders，1868／1969）に基づき，のちにスタンバーグパラダイムと呼ばれる記憶課題を考案した（Sterberg，1966，1969）。スタンバーグパラダイムのなかで, Sternberg は Donders の減算法を加算法へと拡張し，反応時間を用い情報処理ステージを調べる方法を提案した。これを機に，反応時間を用いて情報処理のボックスモデルの検討がなされるようになった。情報処理のボックスモデルは心をコンピュータになぞらえてモデル化したもので，工学のアイデアを援用し認知心理学の分野で使われるようになった。1960年代は心理学史におけるいわゆる認知革命の時代であった。この革命の影響を受け，分野や領域を越えてその影響は広がった。結果として反応時間は，認知心理学だけでなく，知覚心理学，社会心理学，人格心理学，異常心理学，産業心理学，臨床心理学など幅広い領域や分野で現在使われている。研究の現場では知覚，記憶，言語，問題解決など人間の種々のパフォーマンスを測定するだけでなく，偏見や知能，人格のような複雑な構成概念の測定にも使用されている。

Donders の時代から，反応時間測定に関する最大の関心はその精度にあった。ただし，測定機器の精度は古くから極めて高かった。1840年イギリスの科学者 Charles Wheatstone は砲弾速度を測定する機器を発明した。これが正確な時間測定への道を開いた。1842年には，この Wheatstone の発明に基づいて，スイスの時計技師 Mathias Hipp が，500Hz で振動する音叉のような機関を内蔵する計時装置を開発し，ヒップ・クロノスコープと名付けた。さまざまな技術者がヒップ・クロノスコープの改良を重ね，19世紀末には 1 ms の精度での時間測定が可能になった（Popplestone & McPherson，1994）。現在では市販のパーソナルコンピュータでこの精度の測定が比較的手軽に行える。人間の反応時間のばらつきを考えると 1 ms の精度で測定ができるなら，ほとんどの心理学実験において測定精度の面では問題がないであろう。それより優れた精度で測定しても，通常は人間の反応のばらつきが遙かに大きいためあまり意味をなさないからだ。

反応時間の分布

このような反応のばらつきこそが，反応時間測定の精度を考える上で最大の問題となってくる。統計学的に見ると，反応時間はランダム変数として扱わなければならない（Luce，1986）。すなわち，同一の被験者が全く同一の条件で測定を行ったとしても，測定された反応時間には変動が生じ，ある程度の範囲でばらつくのである。

受稿日2010年 9 月29日　受理日2010年12月 7 日

1　専修大学人間科学部心理学科（Department of psychology, Senshu University）

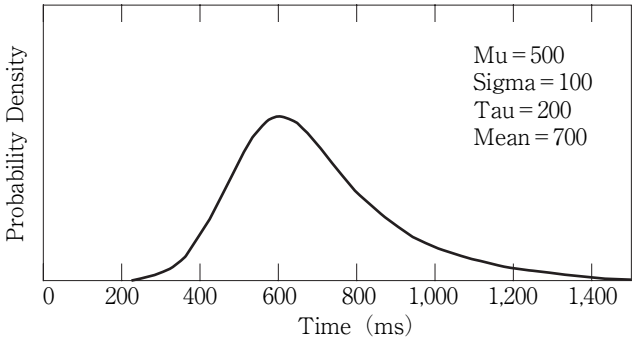

Figure **1.** A ex-Gaussian distribution of reaction times

多くの研究者は反応時間の分布は図１のような形をとると考える（e.g., Luce，1986；Matzke, & Wagenmakers，2009；Ratcliff，1979，1993；Van Zandt，2000）。この分布は，一見，釣り鐘型の正規分布とよく似ている。ただし，正の方向に長く伸びた尾部があり，正規分布と異なり分布の形状は左右非対称である。数学的には，ガウス関数と指数関数の畳み込みをおこなったものとして扱うことが可能である。Luce（1986）はこの形状を指数ガウス関数（ex-Gaussian function）と呼んだ。指数ガウス関数は３つのパラメータで表現できる。それぞれ，μ，σ，τ である。μ はガウス関数の平均値であり，図１で言えば左側にある大きなふくらみの一番高いところの横軸の値におよそ相当する。σ はその標準偏差であり，おなじく左側の大きなふくらみの横幅を決めるものである。一方，τ は正方向の尾部を決定する指数関数のパラメータである。なお，反応時間の分布は，指数ガウス関数だけでなく，Wald 関数，ガンマ関数などでも記述可能である。詳しくは Luce（1986）や Ratcliff（1993）を参照してほしい。ただし，どの分布でも指数ガウス関数の形状のように，大きな山状の分布に付随して正の方向に長い尾部が付くかたちとなっている。

変動をもたらす要因

反応時間は，さまざまな心的状態や測定状況に影響される。年齢，性別などさまざまな個人差だけでなく，疲労や動機づけ，課題への馴れならびにそれに付随する学習効果など刻一刻と変化する個人内の変動も反応時間に影響する。刺激の強さや周囲の雑音などの測定状況も当然影響を与える。もちろん実験における独立変数の操作も影響する。心理学実験における反応時間では，長くとも１s 程度の長さを扱うことがほとんどである。このような短い時間範囲では実験の操作に加え，上に並べた反応時間に影響を与える種々の要因のほぼすべては，加算的にすなわち反応時間を上昇させる方向に作用することになる。そのため，反応時間の分布には正の方向に長い尾部が付くと考えられる。

Luce（1986）は理想的な状態で純粋な反応時間が測定できるなら，視覚刺激に対する単純反応時間の値はおおよそ100 ms なることを示した。この100 ms という時間は視覚系と運動系の生理的過程を反映するものだ。しかし，実際の視覚刺激に対する単純検出反応時間は，NCAA に所属する大学フットボールのスター選手でもせいぜい200 ms くらいである（Eckner et al.，2010）。平均的な大学生なら，260－300 ms 程度であろう。単純検出反応時間ですら，単純な生理的過程以外の要因で圧倒的にその値が上昇する。選択反応時間など複雑な実験事態ではさらに上昇するであろう。実際の反応時間には，実験操作だけではなく，実にさまざまな要因が影響しており，それらすべてを反映した結果が測定されるのである。

反応時間測定における外れ値

平均値など代表値を用いた分析を行うとき，実験操作と必ずしも関係のない影響を受けた値が含まれることは望ましくない。実験の効果が代表値に正確に反映されないからである。特に平均値では，１つの大きな外れ値で，平均値が大きく変わることがある（Ratcliff，1973）。例えば，反応時間の測定を５回行い，（２，２，２，２，12）というデータが得られたとしよう（単位は s）。最後の１回の測定では極めて反応時間が長くなった。これは被験者が疲れて寝てしまったせいだとする（実験ではしばしばあり得ることである）。この場合，平均値は４s となる。最後の回を除いた反応時間はすべて２s なので，最後の１回のために平均反応時間は，１－４回の反応時間とくらべ２倍の値となった。このように原因がはっきりしているとき，外れ値を除くことに多く

のひとはあまり反対しないだろう。代表値をもちいた分析は一般的な傾向を把握するために行われるので，明白に一般的な傾向から逸脱しておりその理由が明らかなものはむしろ除くべきだからだ。

反応時間においても，比較的議論の余地なく外れ値として扱うことが出来るものがある。これは，100 ms 以下の反応時間で焦燥反応と呼ばれる。上述したように，単純検出反応時間に必要な生理的過程のみでもおよそ100 ms が必要される。従って，これよりも速い反応は人間にとって不可能だと考えられる。100 ms 以下の焦燥反応は，実際の刺激に対する反応ではなく，予測に基づくものと判断できるため，「反応」時間の範疇ではない。だからこそ，議論の余地なく外れ値として扱うことが出来る。

一方，外れ値として判断することが必ずしも容易でないものがある。「変動をもたらす要因」の節で議論したようなさまざま要因の影響をうけ，実験操作とは関係のない要因のため増大した反応時間である。反応時間の増大は，実験操作によっても生じうるし，それ以外でも生じうる。純粋に反応時間のみを分析するだけで，それらを切り離すことはとても難しい。

この難しさは，分布の中央付近に位置するものの，不適切な反応に基づいて測定された反応時間について考えると分かりやすいかもしれない。例えば，選択反応時間を測定する実験で，ある被験者があくびをしながら刺激も見ずに反応したとする。この反応時間全体の分布は図１のような形状になるとしよう。そして，たまたま，この「あくびをしながらの反応時間」が600 ms だったと考えよう。この値は図１の分布のおよそ中央である。この数値だけを見て，これが「あくびをしながらの反応時間」であるか適切な選択反応時間であるか判断することは不可能である。残念ながら，実際のところ適切で厳しい統制を行う以外，このような問題に対処することは出来ない。

それでも，繰り返し述べているように，実験操作と関連のない剰余変数による影響は，ほとんどの場合反応時間を上昇させる。従って，実験操作以外の剰余変数の影響を受けた外れ値は，多くの場合，分布の右寄りに大きくずれたものとなる。実際の反応実験では，このような，右寄りに大きくずれた外れ値をどのように扱うかが問題になる。ただし，この右よりの部分にも実験操作を反映した反応時間が存在する可能性は高い（後述の変換アプローチおよびフィッティングアプローチの節で詳しく論ずる）。

さて，今，「大きくずれた外れ値」と述べたが一体いかなる根拠によりそう判断できるのだろうか？何をもってずれたと判断し，何をもってそのずれが大きいと判断するのだろう。一般的に言って，データの逸脱を定量的に判断する先見的かつ客観的な基準はない。外れ値を決める根拠は，研究者がその実験状況を勘案し，得られたデータやこれまでの先行研究などの知見とあわせ，主観的に判断するしかない。

反応時間における外れ値の処理

一般論から言って，外れ値の取り扱いに先見的かつ客観的な基準はない。しかしながら，これまで行われた数多くの反応時間実験の結果を受け，反応時間実験における外れ値の処理方法がいくつか提案されている。本論文では，そのような処理方法のうち代表的なものをいくつか取り上げ，その長所と短所を紹介すると共に，実際の研究場面でどのような処理方法を採用すべきか提言をまとめる。

外れ値処理の種類

反応時間の外れ値の処理には大きく分けて３つある。それぞれ，(1) 選択アプローチ，(2) 変換アプローチ，(3) フィッティングアプローチと呼ぶ。選択アプローチと変換アプローチは比較的古くからあるもので，Whelan (2008) は，これらを中心傾向化アプローチと呼んだ。この２つのアプローチが，平均値や標準偏差のような中心化傾向があるときに有効な代表値を算出することからこのように呼ばれる（Whelan，2008)。一長一短があるものの，研究の現場において最も多く使われるものは，中心傾向化アプローチの一種である選択アプローチである。一方，フィッティングアプローチは比較的新しいものである。なお，フィッティングアプローチは外れ値そのものの処理ではなく，外れ値を含んだ分布全体として分析を行う手法である。また，それぞれのアプローチについても，いくつかの種類があるのでそれらについても，紙幅の許す範囲で紹介する。

選択アプローチ

反応時間における外れ値の取り扱いで，最も多く用いられるのが選択アプローチである。これは分布全体の中から，外れ値を除外して，平均値などの代表値を計算する方法である。反応時間分布全体の中からある範囲の測定結果だけを選択するため選択アプローチと呼ばれる。

これまでの研究において，さまざまな選択方法が提案

されてきた。代表的なものだけでも（1）標準偏差による選択，(2）任意の値による選択，(4）中央値の使用がある。各々の特徴，長所，短所について簡単に紹介しよう。

標準偏差による選択. 標準偏差を基準に用いた外れ値の除去は，最も多くの研究者が用いる手法である（Miller，1994; Whalen，2008）。この方法では，はじめにデータ全体から平均値または中央値そして標準偏差を求め，平均値もしくは中央値にたいし，研究者が決めた標準偏差に基づいて算出された範囲（通常は標準偏差の２－３倍）から外れたものを，外れ値として扱う。伝統的には，平均値から正負の方向に標準偏差の２倍を越えて離れたものを外れ値として扱うことが多かった（e.g., Anscombe，1960；Barnett & Lewis，1978）。標準偏差の２倍という基準は，帰無仮説検定における有意水準を慣習的に５％に設定することに由来するものである。標準正規分布において，平均値（すなわち０）から標準偏差の２倍離れた値を合計すると分布のおよそ５％を占める。このような帰無仮説検定との対応関係から，標準偏差の２倍という基準が用いられてきた。ただし，近年では2.5倍あるいは３倍という基準を用いる研究が増えている。どちらかといえば，３倍を基準に用いる研究が2000年代以降は多数派であろう。

このような範囲の変化にはMiller（1994）のコンピュータシミュレーションが影響を与えた。彼は，標準偏差の２倍，2.5倍，３倍という３種類の外れ値の基準を用い，外れ値が反応時間の平均値に与える影響について検討した。シミュレーションの結果，(1）サンプルサイズにより，外れ値除去の影響が大きく変化すること，(2）標準偏差の３倍という基準が，全体としては外れ値除去の影響が小さいこと，(3）ただし，標準偏差の３倍ではサンプルサイズによる違いが，２倍の時に比べ，大きくなることを報告した。また，具体的な提言として(4）サンプルサイズに違いがあるときは標準偏差を基準にした選択をするべきではないこと，(5）伝統的な標準偏差の２倍という基準では，サンプルサイズの影響が出にくくなるのはサンプル数15以上であり，サンプル数が20を越えると比較的安定することを挙げた。

Miller（1994）の提言において特に注意すべきことはサンプルサイズの違いである。反応時間実験ではしばしば条件間に正答率の差が生ずる。通常の実験では正答反応時間のみが分析の対象となる。従って，正答率に差がある場合には必然的にサンプルサイズの差が大きくなる。また，意図的な注意を操作する実験や学習効果を調べる実験のように実験条件でサンプルサイズが異なる実験デザインも決して少なくない。このような場合には，標準偏差を基準に用いた外れ値の除去は避けた方がよい。特に標準偏差が３倍の時，サンプルサイズの差による悪影響が強く出ることにも留意するべきである。ただし，正答率が極めて高く，天井効果が生じており，サンプルサイズに条件間でほとんど差がない場合には標準偏差の２倍よりも３倍を基準に用いた方が外れ値除去の影響は小さい。近年の研究で標準偏差の３倍という基準が用いられるのは，多くの反応時間実験では正答率が極めて高く設定されており，サンプルサイズの違いによる悪影響が出にくい実験デザインになっていることが関わっているのであろう。

標準偏差による選択は（1）基準が比較的客観的で，(2）帰無仮説検定との対応があり，(3）手続きが簡便という利点がある。一方で，指数ガウス分布をとる反応時間に対して，正規分布（ガウス分布）を想定した基準で外れ値を除去することには根元的な問題がある。反応時間が持つ分布本来の形状を重視すべきで，その特性を考慮した手法を用いるべきだからだ。また，帰無仮説検定との対応は，実際の手続きとしては分かりやすいものの，帰無仮説検定の有意水準自体あくまでも任意のものであることに注意しなければならない。この基準は，決して客観的なものでない。事実，標準偏差の基準は，有意水準以上に研究者の恣意的判断によって，２倍，2.5倍，３倍など異なるものが選択されがちだ。

任意の値による選択. この手法では，分析対象とする反応時間の上限，下限を任意に決定し，その範囲の外にあるものを外れ値として処理する。上述のように，反応時間の下限については多くの研究者の間で合意がとれており（cf. Luce，1986），100 ms に設定する研究が多い。一方，上限については明確な基準がない，多くの研究では1000 ms や1200 ms など先行研究の平均反応時間から考え，明らかに逸脱した値を設定することが一般的である。ただし，何を持って明らかな逸脱と判断するかに明確な根拠はない。筆者が論文の査読者として審査に加わった経験からも，任意の値による選択で外れ値を除去した場合，基準設定の根拠は，多くの査読者が指摘するポイントである，しかも，筆者が見てきた限り，論文の著者から明確な回答があることはほとんどなかった。

任意の値による選択は，比較的最近広まった手法である。詳細は後述するが，Ratcliff（1993）の研究において，任意の値による選択には，種々の外れ値の除去方法を比較において，最も高い検出力を有することが示され

た。おそらく，この研究をきっかけに広く使われるようになったのだろう。

この手法の問題点は上述のように，選択範囲の上限について，明確な基準がないことである。しかもRatcliff（1993）は，設定する上限の値によって検出力が変化することを示した。明確な基準がなく，さらに基準を変化させることで検出力が変わるのでは，適切な基準の設定は困難を極める。

中央値の使用. 中央値の使用も反応時間の外れ値の取り扱いにおいて伝統的に使われてきた手法である。反応時間がとる非対称の分布の代表値として中央値を用いる利点は，古くから多くの研究者が指摘してきた（e.g., Heys，1973；Marascuilo，1971；McCormack & Wright，1964）。例えば，Heys（1973）は，“非対称な分布を分析対象とし代表値を求める場合は，中央値を報告しなければならない（p.235）”と述べた。

しかしながら，現在，反応時間の代表値として中央値を用いることは極めてまれである。中央値の使用は，1980年代後半までは，比較的多くの研究で採用されていた手法であるが，1990年代後半から使用されることは少なくなった。最近，筆者が査読者として関わった論文においても，中央値の使用はしばしば問題視され，審査の過程で統計処理の全面的なやり直しを求められるケースが見られた。

Milller（1988）は，コンピュータシミュレーションを用い，指数ガウス関数状に分布する仮想的反応時間データにおいて，μとσ（ガウス関数のパラメータ）とτ（指数関数のパラメータ），そして，サンプルサイズを操作し，中央値とμの違いに与える影響について検討した。中央値とμの差は，(1) サンプルサイズが小さいほど大きくなり，(2) 分布がゆがむほど増大した。サンプルサイズが小さく分布が大きくゆがむとき，中央値とμに50msを越える差があった。また，全体的には，外れ値を除外しない平均値の方が，中央値よりもμの推定値との差が小さいことも示された。この結果から，Miller（1988）は，(1) 中央値を使うより，外れ値を除去せず平均値を求めるほうが適切な推定ができること，(2) 特にサンプルサイズに差があるとき中央値の使用は不適切であることを提言した。「標準偏差による選択」の項ですでに述べたが，実験条件間におけるサンプルサイズの違いは，正反応だけを分析対象とする典型的な反応時間実験では頻繁に生じうる。これらの点を考慮すると，中央値の使用はやはり避けるべきであろう。

選択アプローチに共通する問題点と反応時間測定への提言. Ulrich and Miller（1994）は，標準偏差による選択，任意の値による選択，中央値の使用の全てにおいて，μとそれぞれの推定値に差が生ずることを示した。すなわち，選択によるアプローチでは，正確にμを推定することは困難であることが示された。もっとも，比較的正確に推定できる条件は存在するので，出来るだけ安定した結果が正確に推定できる実験条件を設定すべきであろう。

現状では，まず，サンプルサイズが異なる実験条件の設定を避け，20サンプル以上を得られるなら，標準偏差の3倍を外れ値とすることが最も簡便な外れ値を処理する手法であろう。実際，現在でも反応時間の分析において最も多く用いられる外れ値の除去方法はこの手法である。ただし，検出力を考えるなら任意の値による選択を行うほうが優れているかもしれない。この手法を用いるためには適切な上限を設定することが必要である。多くの研究者にとってはそれが難題となるだろう。なお，サンプルサイズが異なる場合，選択アプローチの中では任意の値による選択が優れている。ただし，正答率の差によりサンプルサイズが大きく異なる場合には，通常，反応時間よりも正答率が被験者のパフォーマンスを適切に反映する指標になることは忘れるべきではない。

変換アプローチ

ローデータの変換により外れ値の影響を除去したり，データの歪みを補正したりすることが可能である。反応時間のように正の方向に歪んだ分布は，対数変換や逆数変換によって正規分布に近づけることができる（Osborne，2002）。結果として，外れ値の影響を減ずることも，手続き上は可能である。正の方向への歪みについて，対数変換よりも逆数変換に強い補正効果がある。従って，歪みが強いときは逆数変換を用いると補正効果が大きい。また，上述のRatcliff（1993）による比較で，逆数変換は，任意の値による選択に次いで高い検出力を示した。

歪んだ分布から正規分布へ近づける補正は，推測統計を行う上で利点が大きい。反応時間実験の分析には，分散分析やt検定など主にパラメトリック検定が多く使用される。これらの検定では，前提として，検定対象のデータが正規分布していなければならない。歪んだ分布は，対数変換や逆数変換により正規分布へ近づける事が可能なので，このような変換は推測統計の見地から見る限り望ましい。ただし，対数変換や逆数変換によって，分布が必ず正規化するわけではない。この点はしかと留

Table 1. Effects of logarithmic and inverse transformations on variables.

Transformation	x_1	x_2	d_1	y_1	y_2	d_2
Non（Raw data）	1.000	2.000	1.000	11.000	12.000	1.000
Logarithmic	0.000	0.301	0.301	1.041	1.079	0.038
Inverse	1.000	0.500	0.500	0.091	0.083	0.008

Table 2. Effects of logarithmic and inverse transformations on higher order differences and ratios.

Transformation	$\lvert d_1-d_2 \rvert$	d_1/d_2
Non（Raw data）	0.000	1.000
Logarithmic	0.263	7.920
Inverse	0.492	62.500

意すべきである。

なお，実際の研究場面において逆数変換や対数変換が使用されることはほとんどない。なぜなら，これらが非線形変換だからだ。表１にローデータ，対数変換値，逆数変換値について，ローデータが$x_1=1$と$x_1=2$の場合並びに$y_1=11$と$y_1=12$の場合を載せた。それぞれの差であるd_1とd_2を比較すると，ローデータにおける差が変換によって著しく変化していることが分かる。変換の結果，ローデータの値が大きいほどローデータにあった差が小さくなり，その効果は対数変換よりも逆数変換で強調される。また，表２に示したように，それらの差の差あるいは差の比で影響はさらに増大する。

逆数変換や対数変換などの非線形変換は，比率尺度である反応時間を順序尺度へと尺度水準を落とすことで，歪んだ分布から正規分布へ近づける補正を行う（Osborne，2002）。従って，反応時間データにある線形性を利用した実験を行う場合，これらの変換を避けるべきである。例えば，第１節で紹介した単純な加算法に基づくスタンバーグパラダイムでは，対数変換や逆数変換など非線形変換は決して用いるべきではない。メンタルローテーションのように反応時間変化の線形性を検討したい場合も同様である。

また，分散分析において交互作用を調べるデザインでも非線形変換を用いるべきでないだろう。交互作用とは，２要因以上の分散分析において，ある要因の効果が他の要因の水準によって異なることを指す。対数変換や逆数変換を行うと，変換対象の値が大きくなるほどローデータにおける差が小さくなる。そのため，水準間でローデータの値が大きく異なると，変換後にその差を解釈するのは極めて難しい。交互作用は，言い換えると，ある水準における差を別の水準の差と比較することである。すなわち，差の差を比べることだ。表２にローデータ，対数変換値，逆数変換値における差の差の比較を載せた。ローデータでは全くなかった差の差$\lvert d_1-d_2 \rvert$は，対数変換では0.26，逆数変換では0.492となった。このように，非線形変換はローデータになかった差を生み出すことがある。これが逆に働くと，ローデータに存在した差を消し去ることがある。非線形変換された結果の解釈には努めて慎重でなくてはならない。

先に，対数変換や逆数変換の利点として，歪んだ分布を正規分布へ補正することが可能であることを挙げた。しかしながら，検定に合わせて，ローデータを変換することはローデータ本来が持つパタンを必ずゆがめてしまう。実験条件間における差が有意か確認するための統計的仮説検定はローデータに存在する差や効果を確率的に表現するため補足的に行われる手続きである。そもそもは，ローデータの尺度水準や分布，分散の性質などから判断し適切な検定方法を用いるべきだ。ローデータをゆがめ，無理やり前提を満たして検定を行うべきではない。これでは全く持って本末転倒である。前提が満たされない場合には，正規性や等分散を前提としないノンパラメトリックな検定をもちいることも有用な選択肢の一つであろう。ただし，ノンパラメトリック検定でもこれらのふたつの仮定が満たされないときがあり，適用には注意が必要である（Zimmerman，1998）。

最後にもう１点述べておこう。図１に示した正方向に尾部がある反応時間の分布は，外れ値の影響のみで生じるのではない。むしろ，この分布は，人間の情報処理特性を基本的には正しく反映している。例えば，Balota and Spieler（1999）は，語彙決定課題における選択反応時間について検討し，反応時間の指数ガウス分布において，ガウス分布は刺激駆動的な自動処理を，一方，指数分布は概念駆動的な注意的，意図的な処理を反映するというモデルを提唱した。このような分布の成分に注目したモデルは記憶検索や視覚的注意，意思決定などさまざまな分野で存在する（for a review，Matzke & Wagenmakers，2009）。すなわち，反応時間分布における正方向に

伸びた尾部は，ただの外れ値の集合ではなく，人間の情報処理過程の正しく反映したものと考えるべきなのである。人間の情報処理特性を正しく反映した値を変換する必要は全くないであろう。

Ratcliff（1993）による外れ値処理の比較

Ratcliff（1993）は，さまざまな外れ値の処理方法について，シミュレーションをもちい，第１種の過誤と第２種の過誤に与える影響を比較検討した。本論文で紹介した３つの選択アプローチならびに２つの変換アプローチはすべて検討対象であった。

このシミュレーションでは，仮想データに対し，上述した種々の外れ値処理を行い，それぞれに対して分散分析を1000回行った。ただし，標準偏差による選択では，多くの研究で使用される標準偏差２－３倍という基準ではなく，１倍と1.5倍という基準が使用された。すでに述べたように，検出力のもっとも高い手法は，任意の値による選択であり，それに続いたのは逆数変換であった。ただし，任意の値を上昇させると，検出力は低下した。

実際の研究を行う視点から考えて興味深いことは，どの外れ値処理の手法をもちいても，p 値や F 値に有意な変化はなかった。これは外れ値除去の影響が，第１種の過誤には比較的頑健であることを示している。

フィッティングアプローチ

選択アプローチと変換アプローチはどちらも，Whelan（2008）が中心化傾向アプローチと呼ぶもので，反応時間の分布の中から中心化傾向の代表値として平均値や中央値，分散の代表値として標準偏差を求めるものであった。だが，繰り返し述べて来たように反応時間は正規分布をしない。従って，中心化傾向を前提に代表値を算出することは，厳密な意味で適切とは言えない。

フィッティングアプローチでは，指数ガウス関数（e.g., Luce，1986；Ratcliff，1979）や指数 Wald 関数（e.g., Schwartz，2000）などをもちい反応時間データに最尤法あるいは最小自乗法によるフィッティングを行い，反応時間の分布形状を求める（e.g, Van Zandt，2000）。フィッティングアプローチは，外れ値を除去せず分布全体を捉えるのが特徴である。

実際の反応時間は，決して１つの関数で記述できるような情報処理を反映するのではなく，おそらく複数の異なる時間変化を伴う情報処理過程を反映するものであろう。このような考えに基づいて，反応時間の分布全体を分析対象とする研究が近年は増えてきた。先に紹介した Balota and Spieler（1999）などはその典型例である。

Whelan（2008）は，反応時間を中心傾向化アプローチで処理する限り，全く異なる分布パタンを有する反応時間データ間で，平均値では差が見られない可能性を指摘した。さらに，平均値の差が反応時間分布全体のパタンとは矛盾する差を生み出す可能性すら指摘した。彼は，Hervey，Epstein，Curry et al.（2006）の研究を例に挙げ，中心傾向化アプローチとフィッティングアプローチの結果が乖離することを示した。Hervey et al.（2006）は，Conners' Continuous Performance Test と呼ばれる一種の Go/No Go 課題を用い，健常児と ADHD 児の反応時間を比較した。選択アプローチに基づき，平均値を比較したところ，ADHD 児の反応時間は健常児よりも遅くなった。しかしながら，指数ガウス関数を用い反応時間の分布を求めたところ，μ については，選択アプローチによる平均値とは逆に，ADHD 児で健常児より値が小さいこと（反応時間が速いこと）が分かった。一方，指数関数の要素を分析すると，ADHD 児で健常児より著しく遅いことが分かった。Hervey et al.(2006) の結果は，これまで慣習的に用いられてきた選択アプローチに代表される中心化傾向アプローチでは，必ずしも人間の情報処理特性を正しく捉えられない可能性を示唆している。

理論的にも数学的にも，フィッティングアプローチは本論文で紹介した中で，反応時間データが持つ本来の性質をもっとも正確に反映する記述統計手法であろう。ただし，このアプローチには重大な短所がある。その短所のため，多くの研究場面では用いることが困難である。実際，反応時間研究のうち，フィッティングアプローチを用いるのはごくわずかである。

その理由は，反応時間分布を求めるために必要な実験試行数にある。Luce（1986）によれば，典型的な心理学の実験では50－100回のサンプル数から１条件の反応時間の代表値が求められる。Luce（1986）は心理物理学的な背景を持つ研究者なので，この回数はおそらく心理物理学のような，詳細な物理変化を定量的に操作する実験事態におけるものであろう。典型的な認知心理学実験では10－30のサンプル数から代表値が求められる。フィッティングアプローチを用いるときの試行数はこれらの比ではない。Luce（1986）によれば，反応時間分布を得るためには，各条件につきおよそ1000試行が求められる。心理学実験において，１条件のデータのみを求めることはほとんどない。従って，実験条件の数だけ試行

数は倍増する。1試行を4－6秒として1000試行を行うだけでも，1，2時間が必要である。たった1つの条件でこれほどの時間がかかるなら，多数の条件がある実験デザインにおいてフィッティングアプローチを用いることは現実的には極めて困難であろう。Rounder，Lu，Speckman，Sun，and Jiang（2005）のように比較的少ない試行数で反応時間分布をもとめる手法も開発されてきた。それでも，従来の中心化傾向アプローチに比べれば圧倒的多数の試行数が求められることに変わりない。

Hervey et al.（2006）の知見のように，これまで中心化傾向アプローチで分析されてきたデータもフィッティングアプローチを用いることで，新たな側面を明らかに出来るかもしれない。ただし，全ての実験でこのアプローチを用いることは困難である。そのため，既存の実験パラダイムと比較検討し，種々の条件からいくつかを選び出した上で，フィッティングアプローチによる検討を進めるべきであろう。フィッティングアプローチにより，新たな側面が見えてきた場合には，その結果に基づき，さらなる検討を進めるとよい。

結語：実践への提言

本論文では反応時間分析における外れ値の処理について，種々の方法を比較検討した。一般的に言って，データの逸脱を明確に判断する先見的かつ客観的な基準はない。また，外れ値を決める根拠は，研究者がその実験状況を勘案し，得られたデータやこれまでの先行研究などの知見とあわせ，主観的に判断するしかない。これは反応時間の分析でも全く同様である。それでも，代表値をもちいた分析が一般的な傾向を把握するために行われる以上，明白に一般的な傾向から逸脱しておりその理由が明らかなものは分析対象から除くべきである。

本研究では，大きく分けて伝統的な中心化傾向アプローチと比較的新しいフィッティングアプローチを紹介した。フィッティングアプローチは理論的にはもっとも望ましい反応時間の処理方法である。しかし，現実的にこれを行うのは困難である。人間の情報処理は，試行を重ねるたびに刻一刻と変化する。このような性質から考えても，このアプローチの適用範囲はやはり限定的である。また，1000を越える試行数が必要なフィッティングアプローチは現実的な被験者の負担を考えると，実際に用いることは簡単でない。

少なくとも現時点では，伝統的な中心化傾向アプローチとフィッティングアプローチを併用して研究を進めるべきであろう。多くの実験条件を比較したい場合には，中心化傾向アプローチが現実的な制約から考え向いているし，少ない条件を比較し定量的な分析を行いたい場合にはフィッティングアプローチが有効であろう。フィッティングアプローチの結果から，中心化傾向アプローチの結果に疑義がある場合には，それに基づいて，フィッティングアプローチを用いた再検討を行う必要があるだろう。

中心化傾向アプローチでは，条件間でサンプル数の違いがない場合，標準偏差による選択か任意の値による選択を用いるべきである。標準偏差による選択では，バイアスが生じにくい標準偏差の3倍を用いることがよいであろう（Miller，1991）。条件間でサンプル数の違いが大きい場合には，任意の値による選択を行うとよい。ただし，任意の値の上限について，その値を採用した根拠を明らかにする必要がある。最も避けるべき事態は，いくつかの値を試し，実験仮説に適合するパタンを選択することである。このようなやり方を行えば，検定を繰り返す事による有意水準の上昇がおこり，どこかで有意差が生じても決しておかしくはない。

反応時間はさまざまな実験で用いられる反応指標である。知覚，記憶，言語，問題解決など人間の種々のパフォーマンスを測定するだけでなく，偏見や知能までのような複雑な構成概念の測定にも使用される。当然のことであるが，データの測定は出来る限り正確に行わなければならない。その正確なデータは当然適切な統計処理を経て分析されるべきである。反応時間については，その汎用的な利用があるにもかかわらず，必ずしも分析手法が確立されているとは言い難い。今後，フィッティングアプローチのような新しい手法がさらに進歩することで，測定された人間の情報処理をもっと正確に分析できる手法が開発されるであろう。それまで，本論文でまとめたように，中心化傾向アプローチとフィッティングアプローチを併用することが，現実的な取り組みかたであると考える。

引用文献

Anscombe, F.J. (1960). Rejection of outliers, *Technometrics*, **2**, 123–147.

Balota, D.A., & Spieler, D.H. (1999). Word frequency, repetition,and lexicality effects in word recognition tasks : Beyond measures of central tendency. *Journal of Experimental Psychology* : *General*, **128**, 32–55.

Barnett, V. & Lewis, T. (1978). *Outliers in statistical data*. New York : Wiley.

Donders, F.C. (1868). Over de snelheid van psychische proc-

essen.Onderzoekingen gedaan in het Psysiologisch Laboratorium der Utrechtsche Hoogescchool, *Tweede reeks* **2**, 92−120. *Trans*. Donders, F.C. (1969). On the speed of mental processes. *Acta Psychologica*, **30**, 412−431.

Eckner, J.T., Kutcher, J.S., & Richardson, J.K. (2010). Pilot evaluation of a novel clinical test of reaction time in National Collegiate Athletic Association Division I football players. *Journal of Athletic Training* **45**, 327−333.

Hervey, A.S., Epstein, J.N., Curry, J.F., Tonev, S., Arnold, L.E., Conner, C.K. Hinshaw, S.P., Swanson, J.M., & hechtman, L. (2006). *Child Neuropsychology*, **12**, 135−140.

Heys, W. (1973). *Statistics for the social sciences*. New York : Holt, Rinehart, & Winston.

Lachman, R., Lachman, J.L., & Butterfield, E. C., (1979) *Cognitive psychology and information processing*. Hillsdale, NJ : Lawrence Erlbaum Associate

Luce, R.D. (1986). *Response times* : *Their role in inferring elementary mental organization*. New York : Oxford University Press.

Matzke, D., & Wagenmakers, E. (2009). Psychological interpretation of ex−Gaussian and shifted Wald parameters : A diffusion model analysis, *Psychonomic Bulletin & Review*, **16**, 798−817.

Marascuilo, L.A. (1971). Statistical methods for behavioral sciences. New York : McGraw−Hill.

McCormack, P., & Wright, N. (1964). The positive skew observed in reaction time.*Canadian Journal of Psychology*, **18**, 43−51.

Miller, J. (1988). A warning about median reaction time. *Journal of Experimental Psychology* : *Human perception and performance*, **14**, 539−543.

Miller, J. (1994). Reaction time analysis with outlier exclusion : Bias varies with sample size. *Quarterly Journal of Experimental Psychology*, **43 A**, 907−912.

Osborne, J. (2002). Notes on the use of data transformations. *Practical Assessment*, *Research & Evaluation*, 8(6), Available on line [http : //PAREonline.net/getvn.asp?v=8&n=6 September 20th, 2010].

Popplestone, J.A. & McPherson, M.W. (1994) *An illustrated history of American psychology*, Akron, OH : Univ. of Akron Press.

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, **86**, 446−461.

Ratcliff, R. (1993). Methods of dealing with reaction time outliers. *Psychological Bulletin*, **114**, 510−532.

Rouder, J.N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, **12**, 199−223.

Schwartz, W. (2001). The ex−Wald distribution as a descriptive model of response times. *Behavior Research Methods*, *Instruments*, & *Computers*, **33**, 457−469.

Sternberg, S. (1966). High speed scanning in human memory. *Science*, **153**, 652−654.

Sternberg, S. (1969). The discovery of processing stages : Extensions of Donders' method. *Acta Psychologica*, **30**, 276−315.

Ulrich, R., & Miller, J. (1994). Effects of truncation of reaction time analysis. *Journal of Experimental Psychology* : *General*, **123**, 34−80.

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, **7**, 424−465

Whelan, R. (2008). Effective analysis of reaction time data. *Psychological Record*, **58**, 475−482.

Zimmerman, D.W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education*, **67**, 55−68.