岩手大学教育学部研究年報　第67巻（2008．02）81〜94

項目反応理論によるストレス尺度の検討

阿久津　洋　巳＊

（2007年12月７日受理）

１．問題と目的

ストレスは日常よく使われる言葉であり、マスコミで話題になることもある。学生は、宿題と試験、あるいは友人関係によるストレスに悩み、教員は授業と雑務のストレスを同僚や家族にこぼす。健康に関するテレビ番組は、ストレスがいかに心身に有害であるかを説明し、健康に気遣う中高年の人は、時折自分のストレスが高すぎないかと心配する。ストレスは直接間接に心身の健康に影響を与える。例えば、繰り返しストレスを受けると不安障害をおこしたり、慢性的ストレスによって無力感やうつ状態を生じることが知られている。また、ストレスによって脳幹の青斑核の活動が促進されると、睡眠障害を起こす。さらに、ストレスは、新しい神経細胞の成長を阻害することも知られている（具体的なメカニズムとしては、ストレスによって分泌されるホルモンの一種であるグルココルチコイドは、海馬に作用して、神経細胞の新生を助ける brain-derived neurotrophic factor（BDNF）の濃度を低下させる、と疑われている）（Sapolsky，2003）。不必要なストレスを軽減することは、心身の健康な発達とその維持のために重要であろう。

そこで、ストレスの状態を簡易に測定できる用具があると便利である。1 つには、人の生理的反応を測定する方法がある。心拍数の他に唾液に含まれるコーチゾルの量やアミラーゼの量を測定してストレスの程度を推測する方法が知られている（永岑・室田・清水，2002；山口・金森・金丸・水野・吉田博，2001）。しかし、コーチゾルの測定は、簡易とはいえないし、アミラーゼの測定は、簡易に使える商業製品が出回っており興味深いが、測定値はまだ十分検討されていない。

ストレスを心理学的に測定しようとするときは、質問紙法を使うことが多い。すでに発表されているものに、「心理的ストレス反応尺度」（新名・坂田・矢富・本間，1990）、「中学生用学校ストレッサー尺度」（岡安・嶋田・丹羽・森・矢富，1992）、「対人ストレスイベント尺度」（橋本，1997）、「SRS-18」(鈴木・嶋田・三浦・片柳・右馬埜・坂野，1998) その他がある。これらの中で、SRS-18は、項目数が18と少ないため実施が簡単であるばかりでなく、日常に経験される程度のストレス反応を測定できるとされる（鈴木ほか，1998）。SRS-18も含めて、上記のストレス尺度は、どれも古典的テスト理論に基づいて作成された。古典的テスト理論に基づいて作成された尺度にはいくつかの限界があり（Embretson，1996）、新しいテスト理論と呼ばれる項目反応理論を適用して、役に立ちそうな既存のストレス尺度を再検討することが望ま

＊　岩手大学教育学部

しい。

古典的テスト理論の特徴は、テスト得点を項目得点の合計と定義し、目的によっては、それを一次変換（例えばｚ得点）もしくは単調変換（例えば正規化ｚ得点）した得点を測定値として使うことにある。テスト得点は、項目得点の線形結合と見なされる。測定しようとする特性（あるいは能力）をｘ軸に、項目に対する反応の確率をｙ軸に置くと、特定の項目に対する反応確率（yes－no のような二値反応の場合は、yes 反応の確率）は、一般には logistic 関数で近似されることが多い単調増加関数となる。古典的テスト理論が前提とする項目得点の線形結合をこの項目反応曲線の観点から見ると、複数のテスト項目に対応する項目反応曲線があり、その傾きは確率0.5において等しく、項目間の違いはｘ軸方向の平行移動（位置の parameter）のみに現れるという仮定となる。同じ合計得点をもつ２人は、各項目の得点の組み合わせに関係なく、同じ量の特性（あるいは能力）をもつと解釈される。しかし、現実にはテスト項目に対する反応は、項目反応を logistic 関数として描くと、位置の他に傾きの点でも異なることが多い（Rasch モデルが適合しない）。その結果、同じ合計得点が同じ特性の量に結びつかない場合が少なからず生じる。この場合、合計得点と特性の量の間に１対１の対応がないので、合計得点は信頼性には限界がある。この方法上の問題は、しばしば指摘されているが（例えば、Gluck & Spiel，2007）、特性（ストレス反応）の差が比較的小さい個人や集団を比較検討する場面では、特に重要な問題となるであろう。より信頼できる測度は、複数の項目に対する反応のパターンから、テストが測定しようとする特性の量を推定する方法によって得られる。この目的を実現する方法が項目反応理論である。項目反応理論の重要な役割は、（１）項目の測定器具としての性能を調べること、（２）多くの項目に対する反応パターンから、測定していると期待する特性（潜在特性と呼ばれる）の量を推定すること、および、（３）推定の誤差（標準誤差）を計算することである。

長い間、項目反応理論は理論としては理解されていたが、日本では通常、能力テストや性格と心理特性を調べるテストには適用されなかった。項目反応理論を使ってテストを作成するには、特定の高価な統計ソフトウェアーが必要とされることもその理由の一端であろう。最近になって、無償の統計と計算のソフトウェアーである R に項目反応理論を適用したデータ処理の package である ltm が加えられたため、項目反応理論の概略を理解しただけでも、この理論を応用したテストの分析が可能となった（CRAN；Rizopoulos，2006）。

この論文の目的は、ストレス反応を測定する SRS-18を用いて収集されたデータに対して、古典的テスト理論を適用した分析と項目反応理論を適用した分析を実行し、その結果を比較し、項目反応理論の実用的利点を検討することである。

２．方法

（１）質問紙　SRS-18

ストレス反応を調べる SRS-18は18項目から構成されており、３つの下位尺度をもつ：①抑うつと不安、②不機嫌と怒り、③無気力である。比較的軽いストレス反応も測定できると仮定されている。実験状況で繰り返しストレス反応を測定できるように、単純な構造をもつ尺度であり、さらに、ストレスの負荷がかかる状況で生じる一時的なストレス反応を測定できると説明されている（鈴木ほか，1998）。SRS-18を使った予備調査を41名に実施し、項目分析の手順

に従って項目の平均値、合計得点との相関、項目得点の分布および因子分析結果などを検討した。その結果、1つの項目が全項目の合計得点と低い相関を示したため（積率相関係数 $r < 0.2$）、この項目を削除し、残った17項目を使用して質問紙を作成した。付録表1に使用した項目と削除した項目を示す。項目への回答は、1（全くちがう）、2（いくらかそうだ）、3（まあそうだ）、4（その通りだ）の4件法である。

（2）被調査者と実施方法

岩手大学の大学生（工学部、農学部、人文社会学部、教育学部）の664人に、2007年10月の授業時に質問紙を配布し、回答後その場で回収した。

（3）データスクリーニング

欠損値がある被調査者、17項目中16項目以上に同じ反応をした被調査者、学籍番号が書かれていない被調査者（重複が検出できなくなるので）のデータは分析から除外された。さらに、科学的研究の被験者に対する国際的なガイドライン（American Psychological Association, 2001）に沿うように、調査データの使用に同意しなかった被調査者のデータを分析から除外した。同じ被調査者が異なる授業2つ以上で調査を受けた場合、被調査者に重複があった。この場合は、最初の調査データのみを分析対象とした。これらの処置をした後に使用可能なデータとして残った474人の被調査者のデータを分析対象とした。データ分析に使われた被調査者の平均年齢は19.35歳（標準偏差1.21）、男性217人、女性255人（不明2）であった。

（4）分析ソフトウェアー

全ての統計分析と数値計算は、R（2.60版）とその関連 package を使って行われた（CRAN）。

3．結果と考察

（1）古典的テスト理論に従った分析

1）1因子の仮定　　まず、17の項目が1因子のストレス反応を測定していることを、因子分析で確認した。スクリープロットと因子負荷量を調べると、1因子、2因子、3因子のいずれの解釈もできた。作成者たちの分析は3因子を選んでいるが、項目全ての得点を合計してストレスの程度を表すことも可能である、と1因子解釈の妥当性を記している（鈴木ほか，1998）。後の分析が容易なので、ひとまず1因子を採用した。後に4）の項で最終的な分析を行う。

2）項目の平均値　　平均点が極端に低すぎる項目と高すぎる項目は、ストレス反応を適切に測っていない。そのような項目を検出する目的で、全ての項目にわたって平均点を調べた。ここでは、豊田（2002）に習って可能な平均点の最小値と最大値を考え、その両端から間隔の15%以内を、極端に低い（高い）平均点と定義した。具体的には、1.45以下と3.55以上の平均点を極端な値と決めた。17ある項目の最小平均点は1.68、最大平均点は2.57であり、平均は2.02であった。極端に平均点が低い項目と高い項目は見つからなかった（表1参照）。表1に掲載した平均点から、全般に選択肢の1、2の方向に反応の偏りがうかがえる。選択肢1と2の「全く違う」か「いくらかそうだ」を選んだ被調査者が多かった。ストレスをあまり感じていない被調査者が多かったか、質問項目に「難しい」項目が多かったと解釈できる。

3）項目得点と合計得点との積率相関係数　　3番目の項目分析は、項目得点と合計得点の関連である。不適切な項目が少し混ざっているとしても、慎重に内容を吟味して作成した多くの項目は全体としてはストレス反応を測定していると見なせるので、項目の合計得点と項目得点

との相関を調べれば、各項目が適切か否かを判断できる。合計得点と弱い相関しか持たない項目は、その項目が全体の項目で測っているストレス反応を適切に反映せず、ストレス反応の測定には役に立たないであろう（ストレス反応の識別力が低い）。項目に対する反応が順序尺度上にあると仮定し、合計得点を間隔尺度と仮定すれば、polyserial 相関が適当であるが、本研究では広く一般に行われている尺度構成法に従い、項目反応とその合計得点の両方に間隔尺度の仮定を採用してピアソンの積率相関係数を計算した。その際、合計得点に当該項目の得点が含まれていると不当に相関の値が高くなるので、当該の項目を除外して合計得点を計算した。項目16（q16）を除く全ての項目が0.5以上の相関を示し、項目16も相関が特に低いわけではないので（r ＝0.482）、全ての項目が役に立つと判断した（表 1 参照）。どの項目もストレス反応を識別していた（識別力が高い）、と考えられる。

表1．ストレス質問紙項目の3種の統計量

項目	平均点	積率相関係数	因子負荷量
q１	1.83	0.549	0.588
q２	2.07	0.708	0.750
q３	2.57	0.590	0.612
q４	1.77	0.626	0.664
q５	1.86	0.628	0.664
q６	1.68	0.564	0.589
q７	1.89	0.576	0.611
q６	1.84	0.614	0.657
q９	2.06	0.716	0.764
q10	1.89	0.643	0.685
q11	2.32	0.599	0.617
q12	1.84	0.653	0.692
q13	2.44	0.661	0.678
q14	2.15	0.506	0.505
q15	2.01	0.550	0.572
q16	2.05	0.482	0.482
q17	2.13	0.521	0.522
平均	2.02	0.599	0.627

４）積率相関行列を用いた因子分析　　次に質問項目が 1 因子から構成されていることを確認した。因子分析を行い第 1 固有値の寄与率が高く、固有値のスクリープロットでも第 2 固有値以降に大きな寄与率を持つ因子がなければ、質問項目は 1 因子からなると判定できる。そこで、積率相関行列を用い、バリマックス回転を使う因子分析を実行した。 1 因子を指定した場合、第 1 因子の寄与率は38.9％であり、 2 因子を指定した場合は、第 1 因子が27.8％、第 2 因子が19.9％であった。寄与率の大きさから見ると、 1 因子で十分と考えられる。スクリープロットも 1 因子の解釈に反しなかった（スクリープロットは紙面の都合で省略）。 1 因子を指定した場合の、各項目の因子負荷量を表 1 に示す。項目の因子負荷量は、最大0.764（項目 9)、最小0.482（項目16)、平均0.627であり、比較的狭い範囲に集中した。共通因子が各項目に同じような影響を与えており、あえて 2 因子や 3 因子を仮定する必要はない（もっとも、因子数を 1 とするか 3 とするかは、尺度の使用目的による。本研究では、一般的ストレス尺度を考えているため 1 因子とする)。

5）信頼性　Cronbach の α 係数は0.916であった。項目間の相関は、0.152から 0.733の範囲であり、平均0.393、標準偏差0.111であった。項目得点と合計得点との積率相関係数は、0.482から0.716の範囲であり、平均0.599、標準偏差0.067であった（詳細は表 1 参照）。尺度の内的整合性は高いといえる。

6）得点換算表の作成　以上の手続きを経た結果、質問紙に含まれる全ての項目がストレス反応を適切に測定していると判断できた。最後に、尺度構成を完成するために、素点を標準得点（z 変換した値で平均 0 、標準偏差 1 ）に変換するための対応表を作成した。素点と標準得点のほかに標準正規得点（分布を正規化した後に求めた z 値）、正規偏差値（標準正規得点を10倍して50を加えた値）および相対累積度数を表にした（付録表 2 ）。

7）被調査者集団におけるストレス反応の分布　被調査者集団のストレス反応の分布を理解するために、合計得点を項目数で除した平均得点と標準得点（z 値）のヒストグラムを図 1 、要約統計量を表 2 に示す。テスト得点の分布が、中央より左に偏る正の歪度（skewness）をもつことがわかる。ヒストグラムから、項目反応カテゴリーのうち得点 1 と 2 に対応する「全く違う」と「いくらかそうだ」の反応頻度が相対的に多く、「まあそうだ」や「その通りだ」と回答しにくい傾向があった（いわゆる「難しい」項目が多かった）と解釈できる。

図 1 の z 得点のヒストグラムを見ると、テスト全体としてはストレスが高い人を検査する目的に適しているようである。使用したストレス反応質問紙は、臨床的なケースには適しているが、それほどストレスを感じていない人たちの少しのストレス経験の違いを検出する目的には適していないと推測できる。たとえば、今回の被調査者は、一般の大学 1 ～ 2 年生が中心であり、ストレスが低かったのであろう。

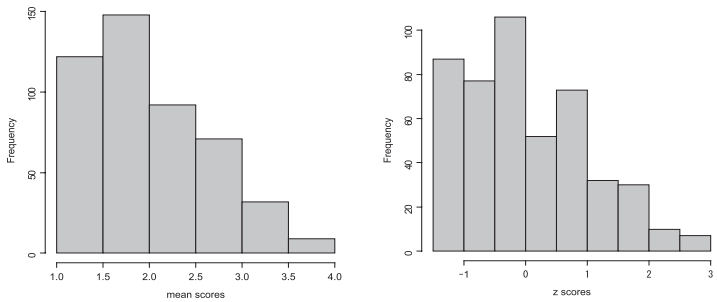

図1．テスト得点のヒストグラム
左の図は平均得点、右の図は標準得点を用いた。明らかに得点の分布に偏りがある。

表2．ストレス反応尺度の要約統計量

	平均得点	標準得点
平均	2.03	0.00
分散	0.42	1.00
中央値	1.88	−0.22
歪度	0.60	0.60
尖度	−0.40	−0.40

（2）項目反応理論に従った分析

古典的テスト理論は、学力測定のテストはもとより、性格特性やストレスのような心理学的特性を測定するために広く使われてきた。心理学的特性の尺度作成を扱う研究が心理学関係の学会誌、たとえば心理学研究、教育心理学研究、パーソナリティ研究などに毎年多く発表されるが、それらの研究は古典的テスト理論に準拠してテストを作成している。本研究が用いた SRS-18も古典的テスト理論に準拠して作成されている。この風潮は、項目反応理論の重要性と有用性が、心理測定の専門家の間で広く知られるようになって久しいのであるから残念なことである（Reckase, 1996）。項目反応理論に従った分析に要する計算ツールは、広く利用され

ている SPSS のような統計ソフトウェアには標準装備されていない。共分散構造分析（SEM）のツールを使うか、MPlus もしくは Mulitlog のような特殊なソフトウェアを使う必要があった（豊田，2002；藤森，2002）。幸いこの計算の技術的な問題は、統計ソフトウェア R に項目反応理論に関する分析ツールが多数含まれることにより解決されつつある（例えば、最近の *Journal of Statistical Software*（May，2007）は、Psychometrics in R という特集号を組んでいるが、掲載論文の大部分は、項目反応理論の新たな展開とそれを実行するソフトウェア（R の package）の紹介である）。

筆者にとって項目反応理論の魅力ある特徴はいくつかある。それらは、(1) 合計得点を使わずに、尺度値を求められる。(2) テストの特性と被調査者の特性を分離でき、被調査者の標本集団に依存しないテストの特性を決めることができる。(3) 一部の項目だけを使用して、信頼できる質問紙を作成できる、(4) 被調査者の特性の推定値（尺度値）に加えてその標準誤差を得られる、などである。しかし、心理学専攻の大学生が卒業研究を行う際に質問紙をしばしば使うが、このような応用研究の際にも項目反応理論に従ったデータの分析が必要なのかは確かではない。

1）項目の選択と 1 因子の確認　　古典的テスト理論の項目分析と同様に、全項目のうちストレス反応とは関係が弱い項目を取り除く必要があり、その目的に polyserial 相関や polychoric 相関を使うことができる。項目に対する反応を間隔尺度ではなく、順序尺度であると仮定し、その合計得点を間隔尺度であると仮定すれば、polyserial 相関を使い、項目反応とその合計得点の両方を順序尺度と仮定すれば、polychoric 相関を使ことが適当であろう（Nunnally & Bernstein，1994）。いずれにせよ、厳密に考えれば、項目反応は順序尺度上にあり、一般に行われている間隔尺度を前提とした処理（積率相関係数）は、理論的難点をふくむ。しかし、他方、項目反応を順序尺度と仮定するならば、項目反応を合計する操作の妥当性が疑われる。そのため、結局は、合計得点を使う polyerisal 相関と polychoric 相関はともに理論的欠陥を含むのではないだろうか。

しかしながら、現実には相関係数を項目選択の大まかな基準として使うのであり、項目反応と合計得点との相関係数自体が、最終的分析に決定的な影響を及ぼすことは少ないと考えられる。このような緩い制限のもとでは、上記 3 種の相関係数（polyserial 相関係数、polychoric 相関係数、積率相関係数）のいずれを使っても実質的な違いはない。どの相関係数を選んでも大きな違いがないことを明確にするために、項目ごとに合計得点との積率相関係数（表 2 と同じ）と polychoric 相関係数を表 3 に示す。たとえば、測定する特性と関連が弱い項目を削除するために、大まかに相関係数0.5を基準とするならば、積率相関係数を使うと16番目の項目（q16）が削除される。polychoric 相関は、積率相関係数より若干大きめの相関係数となり、同じ除外基準を適用すると、全ての項目が排除されずに残る。しかし、0.5の相関は相当に高い（厳しい）基準であり、0.4や0.3の基準を使うこともできる。その場合は、いずれの相関係数を用いても、全項目が基準を満たしている。以下の分析では、全項目を使い項目反応理論を適用して尺度構成を行うことにした。

全項目を使い積率相関行列に基づいた因子分析を行った場合は、すでに 1 因子が適当であるとわかっているので（古典的テスト理論による分析の項参照）、今回は全項目を使い polychoric 相関行列に基づいた因子分析を行った（バリマックス回転）。項目の因子負荷量は、最大0.808（項目 9 ）、最小0.535（項目16）、平均0.679であり、比較的狭い範囲に集中した（表 3 参照）。

表3．各ストレス質問項目と合計得点の相関係数、及び因子負荷量

項目	積率相関係数	polychoric 相関係数	polychoric 相関を使った因子負荷量
q1	0.549	0.661	0.632
q2	0.708	0.793	0.802
q3	0.590	0.701	0.660
q4	0.626	0.736	0.718
q5	0.628	0.741	0.725
q6	0.564	0.688	0.676
q7	0.576	0.672	0.666
q8	0.614	0.713	0.709
q9	0.716	0.790	0.808
q10	0.643	0.743	0.732
q11	0.599	0.717	0.662
q12	0.653	0.753	0.747
q13	0.661	0.761	0.725
q14	0.506	0.617	0.555
q15	0.550	0.674	0.637
q16	0.482	0.606	0.535
q17	0.521	0.623	0.558
平均	0.599	0.705	0.679

積率相関係数を使った因子分析に比べて、やや高めの因子負荷量が得られたが、全体に結果は積率相関係数を使った場合に似ており、1因子構造の解釈ができる（スクリープロットも積率相関係数を使った場合に似ていた)。次に項目母数を推定した。

2）項目母数（パラメータ）の推定　　ストレス反応を調べる質問項目は、4件法で回答を得た（1～4の1つを選ぶ方法をとった)。これらの数値は、程度の順序関係をもつ（1＜2＜3＜4)。本研究では、このような段階反応のデータを分析するのに適したコンピュータプログラムの grm を使用した（gm は、R の ltm パッケージにある（Rizopoulos，2006))。

項目パラメータの1つが、項目の識別力である。項目間で識別力が等しいとは考えにくいが、これは経験的に解決できる問題なので、次のようにデータを分析して確認した。識別力一定の設定と異なる設定の2通りで、grm を使って項目反応分析を行い、どちらのモデルがよりデータに適合するか判定するために、likelihood ratio test（尤度比検定）を実施した。その結果、項目間で識別力が異なる設定の方が有意にモデルのあてはまりがよかったため（尤度比＝102.14，(df ＝16)，p ＜0.001)、識別力を項目ごとに推定した分析結果を報告する。

まず、項目反応カテゴリー特性曲線（IRCCC, item response category characteristic curve）を潜在特性（ストレス反応）に沿って位置づける位置パラメータ（3つ）と識別力パラメータを表4に示す。

識別力に関しては、全項目が識別力1.0以上であるから、どの項目も識別力は十分高かった。しかし、詳しく見ると、項目間で多少ばらつきがあり、反応カテゴリ間の反応確率が比較的明瞭に分かれている項目（識別力大）とその反応確率が明瞭に分かれていない項目（識別力小）があった。例えば、2番目の項目（q 2）「悲しい気分だ」と9番目の項目（q 9）「気持ちが沈んでいる」は高い識別力を示し、反対に、項目16（q16）の「根気がない」は低い識別

力を示した。項目反応理論を適用して求めたテスト得点と「悲しい気分だ」（q 2）と「気持ちが沈んでいる」（q 9）との相関（polyserial 相関係数）は、それぞれ0.847、0.859と高い相関を示した。一方、「根気がない」に対する回答と項目反応理論によって得られたテスト得点のpolyserial 相関係数は、0.556と中程度の相関であった。項目 2 および項目 9 は、テスト得点に対する影響が大きく、反対に項目16は、テスト得点に対する影響力が比較的小さいことが確認できる。

表4．ストレス質問項目の位置パラメータ（b1, b2, b3）と識別力

項目	b1	b2	b3	識別力
q 1	−0.213	1.260	2.535	1.353
q 2	−0.452	0.630	1.547	2.257
q 3	−1.514	−0.029	1.227	1.527
q 4	−0.006	1.241	2.189	1.747
q 5	−0.034	0.906	1.852	1.793
q 6	0.251	1.553	2.424	1.501
q 7	−0.115	0.938	2.109	1.467
q 8	−0.145	1.112	2.161	1.661
q 9	−0.501	0.649	1.596	2.306
q10	−0.267	0.960	2.131	1.788
q11	−0.990	0.399	1.475	1.549
q12	−0.061	1.003	1.927	1.832
q13	−1.087	0.136	1.306	1.826
q14	−0.798	0.702	2.181	1.139
q15	−0.289	0.840	1.758	1.353
q16	−0.601	1.051	2.295	1.050
q17	−0.824	0.839	2.187	1.117

b1は最小カテゴリに対する反応確率が50%になる特性の位置
b2は2番目と3番目のカテゴリに対する反応確率が同じになる位置
b3は最大カテゴリに対する反応確率が50%になる特性の位置

位置パラメータに関しては、3 つのパラメータのうち中央の b 2（反応カテゴリー 2 と 3 の境の z 値）を見ると、b 2 の値が項目 3（q 3「何となく心配だ」）を除き全項目で 0 以上（すなわち、反応カテゴリーの 2 と 3 の境が潜在特性の z 値 0 の右側にある）であり、1 に近いかそれを超える項目も多数ある。3 つの位置パラメータの中央である b 2 が、潜在特性の 0 から 1 の間にあるということは、反応全体が潜在特性の軸上で、右側（値の大きい方）にずれていることを示す（一般的ストレス反応の測定という目的からは、潜在特性の 0 近辺にあるのが理想である）。項目 3 は、他の項目に比べると「まあそうだ」の回答が出やすい項目であった。b 2 の値が一番大きいのは、項目 6「感情を抑えられない」であった（z ＝1.55）。「まあそうだ」の回答が出にくい項目である。この項目に「まあそうだ」や「その通りだ」と答える人は、一般大学生の中には少ないであろう。この項目に「まあそうだ」と回答した被調査者は、48人で全体の10.13%、「その通りだ」と回答した被調査者は、32人で全体の6.75%であった（すなわち、4 件法で 3 か 4 を選んだ人は全体の約17%以下であった）。これらの学生は、何らかの理由で調査時に普段より高いストレス状態にあったか、持続的なストレス状態にあると推測できる。このストレス検査のほとんど全ての質問が「難しい」傾向（ストレス反応を示しにくい傾向）があり、ストレス反応が高い人に適した項目が多いようである。

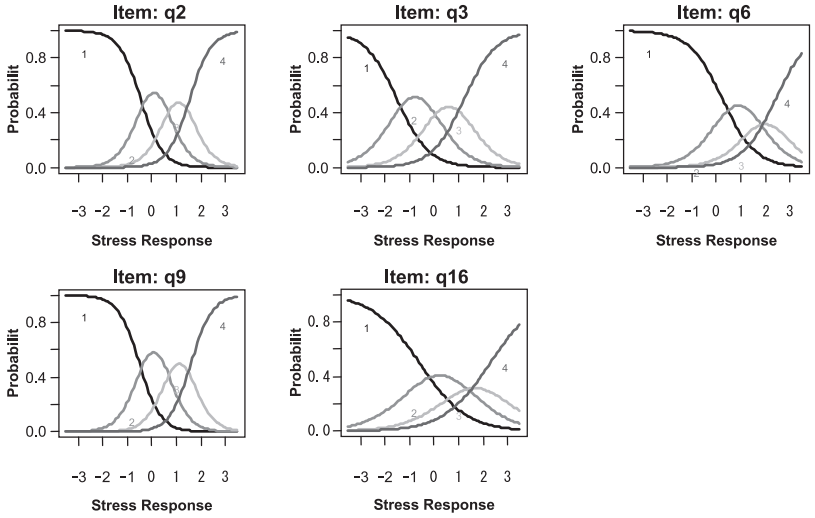

図２．項目反応カテゴリー特性曲線（IRCCC）を識別力が高い項目（q2, q9）と識別力が低い項目（q16）、及び易しい項目（q3）と難しい項目（q6）について示す。

上で考察した項目（項目２、３、６、９、16）の項目反応カテゴリ特性曲線（Item Response Category Characteristic Curve, IRCCC）を図２に示す。項目２と９では、各カテゴリに対する４つのカーブが急峻で、反応の弁別が高い（識別力が高い）。項目16では、４つのカーブがなだらかで、反応の弁別が低い（識別力が低い）。易しい項目（q ３）と難しい項目（q ６）を比較すると、難しい項目ｑ６のカーブは、易しい項目ｑ３のカーブに比べ Stress Response と名前をつけられたストレス反応特性の標準得点（z 値）に対して右に偏っていることに気づく。この図から平均的なストレスを感じている人（Stress Response の軸上の０近辺に位置する人）は、項目６に「全くちがう」か「いくらかそうだ」と回答すると推測できる。残念なことに、項目６（q ６）と対照的に程度が弱いストレスを感じている人が、「そのとおりだ」と回答する項目は見つからなかった。それでも、項目３に対しては、平均的なストレスを感じている人が「まあそうだ」や「その通りだ」と回答する確率は、「全くちがう」や「いくらかそうだ」と回答する確率と同じくらい大きい。項目３のように平均的なストレスを感じている人や弱いストレスを感じている人が、「そのとおりだ」と回答する項目を全項目数の半分程度含めば、このストレス反応検査の適用範囲は広がるであろう。

最後に、尺度値とその推定誤差に関して述べる。（古典的テスト理論の信頼性係数から測定の誤差を推定することはできないが、項目反応理論を適用して得られるテスト情報量（フィッシャー情報量）から測定誤差を推定することができる。フィッシャー情報量の平方根の逆数が尺度推定値の標準誤差に対応する。）本研究で使用したストレス検査のテスト全体の情報量を調べると、全ての得点範囲にわたる情報量が58.05であり、得点が（－４～０）の間に19.18、（０～４）の間に37.47含まれていた。他方、（－４～０）の得点範囲に含まれていた情報量は、全体の情報の33.04％であった。得点が高い領域（０～４の間）に比較的多くの情報が含まれている。テスト情報曲線（図３）はこの傾向を明瞭に示し、テスト得点が－1.0から2.5の範囲で情報量が大きく、この範囲のストレスがある人に今回のストレス反応検査を実施すると、テ

ストの精度がよく、測定は信頼できることがわかる。テスト得点が 0 の被験者に使用しても測定精度が高いことに注目すべきである。目安となる特性値の標準誤差を表 5 に示した。ストレス反応値が－0.5から2.5の間では、標準誤差は0.3以下である。標準誤差の1.96×$\sqrt{2}$倍（大まかに 3 倍）が 5 %の有意水準に相当するから、得られたストレス反応の得点（z 値）が－0.5から2.5の間にある 2 人を比較する場合、ストレス反応がおよそ0.9（0.3×3）異なれば、ストレス反応の程度が異なるといえる（ 5 %水準で統計的に有意差がある）。項目反応理論を適用して求めたテスト得点の被調査者集団における分布をヒストグラムとして図 4 に示す。要約統計量は、［平均0.05、分散1.00、中央値－0.08、歪度0.38、尖度－0.31］であった。歪度が0.38と通常の方法で得点を求めた場合の歪度（0.6）の60%ほどに低下して、正規分布に近づいていることがわかる。ヒストグラムも同様の傾向を視覚的に示している（図 2 と 4 のヒストグラムを比較）。繰り返しになるが、正規分布を仮定して位置パラメータを推定したにもかかわらず、このような歪みが残っているところに、このストレス尺度の不完全さがうかがえる。

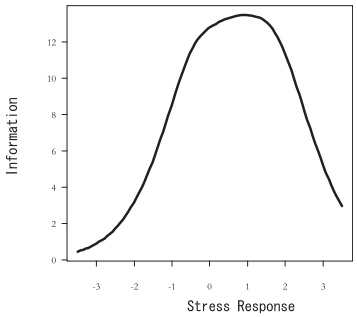

図３．テスト情報曲線
ストレス反応の標準得点（z）の関数としてテスト情報量を調べると、－１～2.5の範囲で情報量が多いことがわかる。

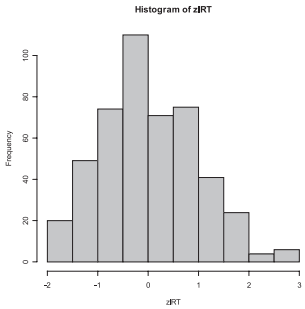

図４．項目反応理論を適用したテスト得点のヒストグラム

表5．ストレス反応得点の標準誤差

−3.0	1.070
−2.5	0.755
−2.0	0.570
−1.5	0.430
−1.0	0.345
−0.5	0.295
0.0	0.280
0.5	0.274
1.0	0.273
1.5	0.276
2.0	0.300
2.5	0.247
3.0	0.445

4．総合考察

テストの合計得点と項目反応理論によるテスト得点の比較

通常テストの得点として計算される合計得点と項目反応理論を適用して得られるテスト得点に実質的な違いはあるのだろうか。項目反応理論によるテスト得点は、テストの合計得点からは直接求められない。テスト得点は、ある特性（観察できない潜在特性）に関して、分析に含まれる全項目に対する被検査者の反応パターンが得られる確率が最も高い特性値として定義される。これら 2 種類のテスト得点の違いを観察するために、古典的テスト理論の手順に従って求めたテストの合計得点と項目反応理論によるテスト得点を比較した（図 5 参照）。図 5 から項目反応理論によるテスト得点（zIRT）はほぼ合計得点に対応しているが、詳細に検討すると同じ合計得点に対応して多数の zIRT があることがわかる。

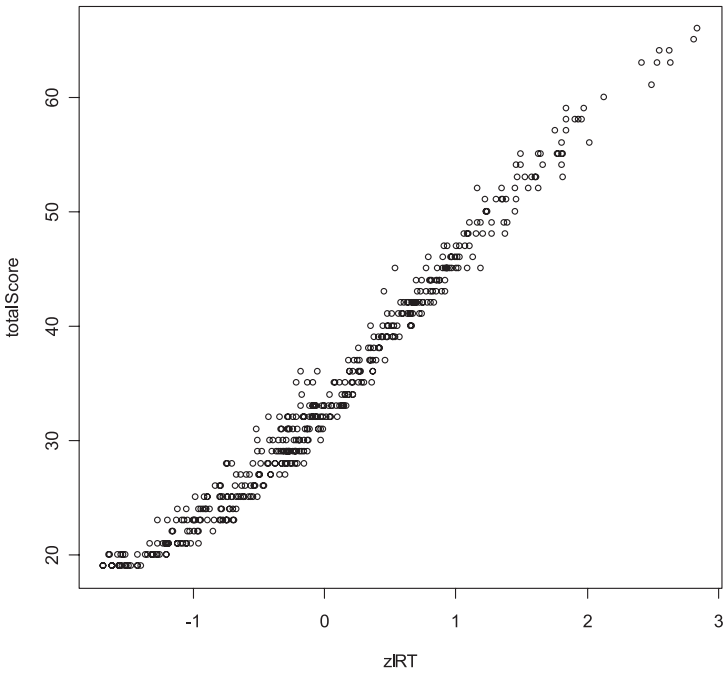

図５．項目反応理論によるテスト得点と合計得点の比較
横軸は項目反応理論によるテスト得点、縦軸は項目の合計得点を表す。

古典的テスト理論では、項目の合計得点を作る手続きがとられるが、合計得点が測定値として厳密に信頼できるのは、上述したように Rasch モデルがあてはまる場合であり、これが当てはまらないときは、合計得点は特性や能力の指標として信頼性に限界がある。本研究のストレス反応尺度のように評定反応（１、２、３、４などのような）のデータでは、Rasch モデルに相当する部分は、識別力を項目間で一定とする仮定である。識別力が項目間で一定ならば、２値反応データに対する Rasch モデルにおいて全ての項目にわたって１つの識別力を適用できる場合と同じである。上述の分析でふれたように、使用した質問項目の間で識別力は異なっていた。したがって、合計得点もしくはその１次変換である z 値（さらに正規化した z 値も）を被調査者のストレス反応の測度として用いることは理論的には好ましくない、という見方ができる。項目反応理論は、合計得点は求めず、全項目の情報を利用して被調査者個人の測度を算出するため、このような理論的問題を免れている。

しかしながら、実用上の観点から図５を検討すると、項目反応理論により得られたテスト得点（zIRT）と項目の合計得点間の違いは、推定誤差の範囲内と見える。同じ合計得点に対する zIRT の広がりは、zIRT の尺度で0.9以内にほぼ収まっている（0.9は標準誤差の３倍、上記のテスト情報量の項参照）。実用上は、項目の合計得点に基づくテスト得点と項目反応理論に基づくテスト得点のどちらのテスト得点を使用しても、大きな違いはないかもしれない。

ただし、注意すべきことは、本研究で使用したストレス反応尺度は、その項目特性が類似していた。合計得点の使用にこのような制限が必要であるかは、項目特性にばらつきがあるテストについて検討した後に結論を出すべきであろう。

項目反応理論の有効な適用範囲として、すでに、項目パラメータが確定している質問紙を使って少数の被験者を検査する状況が考えられる。例えば、被験者に実験的にストレスを経験させ、ストレス反応の程度を調べると同時に別の変数（negative 感情など）を調べるような実験である。このような場合、実験前とストレス経験後にストレス反応を調べ、その値を個人内で比較する際には、項目反応理論を適用して得られる標準誤差は貴重な情報を与える。（通常の分析では、ストレス反応の標準得点（z 値）は得られるが、１人の被験者内の標準誤差を得ることはできない。）標準誤差を利用して、ストレス経験前と後で、ストレス反応に違いがあるか確認できる。必要ならば、もとの質問紙から異なる項目を選び出し、２つの別な質問紙を作成して、その１つをストレス経験前に別な１つをストレス経験後に実施することも可能である。これにより、同じ質問紙の繰り返しを避けることができる。

5．まとめ

本研究は、既存のストレス反応尺度によって得られたデータについて、古典的テスト理論にしたがった分析と項目反応理論にしたがった分析を行い、その結果を比較検討した。まず第一に、ストレス反応テストに含まれる項目の分析においては、項目反応理論による分析は項目の詳細な特性を調べることができた。古典的テスト理論による分析は、この点では十分ではなく、テストを作成する段階およびその特性を検討する際に、項目反応理論の利点は大きい。ついで第二に、項目反応理論にもとづいて計算されたテスト得点と古典的テスト理論にもとづいて計算されたテスト得点（項目の合計得点とそのｚ変換値）には多少のずれがあったが、このずれは得点のおよそ95％信頼区間内に収まった。そのため、少数の被験者を使った事例研究および実験研究などの場合は別として、古典的テスト理論にもとづいて計算されたテスト得点は実用上十分役に立つと考えられる。

謝辞　本研究にあたり、ストレス反応尺度を用いたデータの収集およびデータの入力をしてくれた岩手大学教育学部学校教育教員養成課程小学校教育コース心理学サブコースの浅野壮志さん、小田島裕美さん、宮聡美さんに感謝します。

引用文献

1．American Psychological Association (2001) Appendix C: Ethical standards for the reporting and publishing of scientific information. In *APA Publication Manual*. American Psychological Association (APA). Washington, DC.

2．CRAN The Comprehensive R Archive Network. http://cran.r-project.org/
(This server is hosted by the Department of Statistics and Mathematics of the WU Wien.)

3．Embretson, S.E. (1996) The new rules of measurement. *Psychological Assessment, 8*, 341−349.

4．藤森進（2002）項目反応理論による多値データの分析について　―段階得点モデルと部分得点モデル―『人間科学研究』文教大学人間科学部，24，21−31.

5．Gluck, J. Spiel, C. (2007) Using Item Response Models to Analyze Change. In *Oxford Handbook of Methods in Positive Psychology*. (Ong & van Dulmen, ed)

6．Nunnally, J.C. Bernstein, I.H. (1994) *Psychometric Theory (third edition)*. McGrow-Hill, Inc. Yew York.
7．永岑光恵・室田真男・清水康敬（2002）暗算課題遂行中における唾液中コルチゾールと心拍数を用いた心理変数の評価　電子情報通信学会　信学技法，157–164.
8．新名理恵・坂田成輝・矢富直美・本間昭（1990）心理的ストレス反応尺度の開発．心身医学，30，29–37.
9．岡安孝弘・嶋田洋徳・丹羽洋子・森俊夫・矢富直美（1992）中学生の学校ストレッサーの評価とストレス反応との関係　心理学研究，63，310–318.
10. Reckase, M.D. (1996) Test Construction in the 1990s: Recent Approaches Every Psychologist Should Know. *Psychological Assessment, 8*, 354–359.
11. Rizopoulos, D. (2006) ltm: An R package for Latent Variable Modeling and Item Response Theory Analysis. *Journal of Statistical Software, 17*, 1–25.
12. Sapolsky, R. (2003) Taming Tress. *Scientific American, September*, 67–75.（日本語訳「不安とうつを克服する」日経サイエンス，2003年12月号，80–90.）
13. 鈴木伸一・嶋田洋徳・三浦正江・片柳弘司・右馬埜力也・坂野雄二（1998）新しい心理的ストレス反応尺度（SRS-18）の開発と信頼性・妥当性の検討．行動医学研究，4，22–29.
14. 豊田秀樹（2002）項目反応理論［入門編］―テストと測定の科学―　朝倉書店
15. 山口昌樹・金森貴裕・金丸正史・水野康文・吉田博（2001）唾液アミラーゼ活性はストレス推定の指標になり得るか．医用電子と生体工学，39–3，234–239.

付録

付録表1．ストレス反応検査項目

q1	怒りっぽくなる
q2	悲しい気分だ
q3	何となく心配だ
q4	怒りを感じる
q5	泣きたい気持ちだ
q6	感情を抑えられない
q7	くやしい思いがする
q8	不愉快だ
q9	気持ちが沈んでいる
q10	いらいらする
q11	いろいろなことに自信がない
q12	何もかもいやだと思う
q13	よくないことを考える
q14	話や行動がまとまらない
q15	なぐさめて欲しい
q16	根気がない
q17	何かに集中できない
削除項目	ひとりでいたい気分だ

付録表2．テスト得点換算表

素点	標準得点	標準正規得点	正規偏差値	相対累積度数
66	2.88	3.07	80.7	1.000
65	2.79	2.73	77.3	0.998
64	2.70	2.49	74.9	0.996
63	2.61	2.27	72.7	0.992
(62)	(2.52)	(2.21)	−72.0	(0.989)
61	2.43	2.15	71.5	0.985
60	2.33	2.10	71.0	0.983
59	2.24	2.03	70.3	0.981
58	2.15	1.92	69.2	0.977
57	2.06	1.83	68.3	0.968
56	1.97	1.77	67.7	0.964
55	1.88	1.67	66.7	0.960
54	1.79	1.56	65.6	0.945
53	1.70	1.48	64.8	0.937
52	1.60	1.40	64.0	0.924
51	1.51	1.32	63.2	0.914
50	1.42	1.26	62.6	0.901
49	1.33	1.21	62.1	0.892
48	1.24	1.14	61.4	0.880
47	1.15	1.07	60.7	0.865
46	1.06	1.01	60.1	0.852
45	0.97	0.91	59.1	0.833
44	0.87	0.82	58.2	0.804
43	0.78	0.75	57.5	0.783
42	0.69	0.66	56.6	0.764
41	0.60	0.57	55.7	0.728
40	0.51	0.50	55.0	0.700
39	0.42	0.43	54.3	0.679
38	0.33	0.38	53.8	0.654
37	0.24	0.33	53.3	0.639
36	0.14	0.27	52.7	0.618
35	0.05	0.21	52.1	0.597
34	−0.04	0.15	51.5	0.570
33	−0.13	0.07	50.7	0.551
32	−0.22	−0.04	49.6	0.508
31	−0.31	−0.14	48.6	0.460
30	−0.40	−0.23	47.7	0.426
29	−0.49	−0.33	46.7	0.395
28	−0.59	−0.45	45.5	0.346
27	−0.68	−0.54	44.6	0.306
26	−0.77	−0.61	43.9	0.283
25	−0.86	−0.72	42.8	0.257
24	−0.95	−0.85	41.5	0.213
23	−1.04	−0.99	40.1	0.184
22	−1.13	−1.12	38.8	0.139
21	−1.22	−1.27	37.3	0.122
20	−1.32	−1.54	34.6	0.082
19	−1.41	−2.05	29.5	0.040

( ) 内の数字はデータにはなかったが、可能な値である。