【研究論文】

肯定的項目と否定的項目の混在が尺度に及ぼす影響：項目反応理論による社会調査データの分析

Evaluating the Influence of Mixing Positively and Negatively Expressed Items in a Scale: Social Survey Data Analysis Utilizing Item Response Theory

澁谷泰秀[1]・渡部諭[2]・吉村治正[3]・小久保温[4]

[1]青森大学社会学部, [2]秋田県立大学総合科学教育研究センター, [3]奈良大学社会学部,
[4]青森大学ソフトウェア情報学部

Abstract

The purpose of the study is to evaluate the effects of mixing positively and negatively expressed items in scales on validity and reliability of the scale. Two social surveys conducted in 2011 and 2013 included happiness scales with the same item contents, but slightly different item expressions. The first survey had a scale with all five items expressed negatively and the scale used in the second survey had two items expressed negatively and three items positively. We used item response theory (graded response model) for analyzing those two scales. We reversed all negatively expressed items to adjust the directionality towards happiness. The happiness scale with all negatively expressed items had higher mean scale score than the mixed scale. The scale with all negatively expressed items extracted more information as a scale, although extracted information was mainly from a lower range of the scale scores. The scale with mixed items extracted less amount of information compare to the all negatively expressed item scale, but it extracted information in a wider range of the scale scores. Mixing positively and negatively expressed items in a scale may be beneficial if one could construct a scale in a way that it extracts more information in a wider range of scale scores through applications of item response theory.

***Keywords**: Item wording, positively worded items, negatively worded items, item response theory, graded response model*

１．はじめに

本研究は，社会調査などで用いられる調査票の項目ワーディングが項目反応に及ぼす影響及びその結果として尺度分析や重回帰分析などの統計分析に及ぼす影響の評価を試みた澁谷・渡部・吉村・小久保　(2014) の第２報である．第１報である澁谷・渡部・吉村・小久保　(2014) は，尺度分析に古典的テスト理論を用いた研究であったが，本研究では，尺度分析に項目反応理論を用い，古典的テスト理論と項目反応理論を用いた分析結果の比較を行った．

社会調査の中で最も頻繁に用いられる方法の一つに質問紙法がある．質問紙法は，複数の項目や尺度で構成される調査票を用いて回答者に関する情報や意見などを尋ねる方法である．回答者は，調査項目を解釈し，項目ごとに意思決定をして回答する．質問票を用いて行

われる社会調査は，ほとんどの回答者が質問票の項目を同一に解釈すると前提されている．この前提は，社会調査において調査対象者の属性や具体的な行動を問う調査項目においては，問題となることは少ないが，態度や意見などの心理学的構成概念を測定する抽象的な尺度項目を用いる場合には，内容が同一な項目であっても項目の表現が異なると，上記の前提が自明とはならない場合がある．その為，社会調査の専門家は，調査項目の構築や調査票のデザインは，妥当性及び信頼性を確保するために慎重に行う必要があるしている(Dillman, D. A, 1978; Biemer, P. P. & Lyberg, L. E., 2003；Liez, P. P , 2010)．尺度の中に肯定的な表現を用いた項目と否定的な表現を用いた項目を混在させることは，一辺倒な回答パターンに慣れて項目内容を熟読せずに前項と同様に回答するなどの応答バイアス（response bias）を避けるために行われてきた（Anastasi & Urbina, 1983; Spector, 1992;Nunnally & Bernstein, 1994;）．肯定的項目と否定的項目を同一尺度内に混在させることが，回答者の応答に影響を及ぼさないとする根拠の一つに期待効用理論がある（von Neumann & Morgenstern, 1947）．期待効用理論によると，我々の意思決定（判断）は，意思決定の対象となる事象の効用の期待値に基づいて決定される．効用の期待値は，判断をする際の各選択肢の効用とその効用の出現確率との積である．回答者が社会調査項目に回答するプロセスでは，およそこのような計算が回答者の応答プロセスの中で意識的に又は無意識に生起していると捉えることができる．このプロセスにおいては，各選択肢の効用が調査項目で与えられた条件の内容に基づいて決定される．すなわち，項目が肯定的にワーディングされていても否定的にワーディングされていても，内容が同一であれば，判断の基準となる効用に影響を及ぼさないと規定されている．

意思決定研究においては，数理的表現の一意性は，意思決定問題において数理的な内容が同一であれば，項目の表現が異なっていたとしても，一般的に同一の反応が生起すると仮定されていた．この視点は心理学尺度や社会調査項目の分析にも応用され，設問の意味が同一であれば，表現が異なっていても同一の項目であると判断されることが多かった．このような考え方の延長上に，社会調査などで用いられる尺度が肯定的表現の項目と否定的表現の項目で構成されている場合，否定的項目を逆転させて，肯定的項目と加算して尺度得点とする方法が妥当であるとする考え方がある．

Tvasky & Kahneman (1981) は，回答者が項目を解釈して効用の期待値を予測するプロセスについて，期待効用理論とは異なり，項目の内容は同一であっても，項目のフレーム（肯定的あるいは否定的などの項目のワーディング）が異なる場合，回答者の意思決定は異なる現象が多くあることを報告した．Tvasky & Kahneman（1981）の報告は，現在では広く受け入れられており，項目のワーディングが及ぼす意思決定への影響に関する研究が多く行われている．Kahneman は，この報告に始まるフレーミング効果やそれを説明するプロスペクト理論などの意思決定に関する一連の業績が認められ，2002 年にノーベル経済学賞を受賞している．回答者の意思決定プロセスに関するフレーミング効果やプロスペクト理論などについては，本論の第 1 報である澁谷・渡部・吉村・小久保 （2014）において概要を報告しているため，第 1 報を参照されたい．

1．1　項目反応理論

現在，日本の教育評価，心理学，社会科学の諸分野などにおいて広く用いられているテスト理論は，古典的テスト理論と呼ばれるもので，テストなどの得点を計算する際に，正解の得点を加算して総得点とする方法などの理論的背景となっている．古典的テスト理論は，日本においてほぼ唯一のテスト理論として広く用いられてきたが，理論的弱点が広く認められてきたこと，さらに，理論的により優れている項目反応理論がコンピュータ及び分析ソフトなどの発達によって実用化されるようになったことから，現在では大規模テストなどでは使用されることは少なくなった．

項目反応理論（Item Response Theory: IRT），は，欧米において特に 1980 年代以降，マイクロコンピュータ及び項目反応理論関連アプリケーションソフトの発達に伴い，広く用いられるようになってきた．項目反応理論では，回答者の項目への反応と測定する特性との関係を関数としてモデル化し，そのモデルに基づき，各回答者の項目反応パターンを最も高い確率で創出する特性値（能力推計値）を推計する．推計された特性値は，その特性値を導く尺度項目の特性に影響を受けない事，及びその尺度項目の特性（難易度パラメタや識別力パラメタ）が回答者の能力などの特性に左右されないこと，などが古典的テスト理論と比較して優れている点とされている．

項目反応理論の古典的テスト理論に対する理論的優位性は確立しているが，変数の特性値（Scale Score）や項目特性パラメタの推計には比較的複雑な計算を必要とする為，BILOG, LOGIST, EQSIRT, IRTPRO などのコンピュータプログラムが必要である．また，古典的テスト理論のように感覚的な理解が困難で，分析に慣れるまでは計量心理学者などの専門家の助力が必要であることなどから，日本においてはまだ一般的認知度は低い．

項目反応理論を用いた分析では，使用された項目反応モデルと回答者の項目反応パターンの適合性 （model fit） を評価する必要があり，使用された項目反応モデルとデータとの適合性が低い場合には，項目反応理論は有効な分析手法とはならない．

現在，最も頻繁に使用されている項目反応理論モデルの一つに 2 母数ロジスティックモデルがある．この項目反応モデルは，各項目に対する反応確率が 2 つの項目母数（項目の難易度及び項目の識別力）と被験者の能力推計値（スケールスコア）との関数で表現できるとするもので，項目反応確率は式（1）で表現される．

$$P_i\,(\,x = 1 \mid \theta_j) \;=\; \frac{1}{1 \;+\; e^{-a_i\,(\theta_j - b_i)}} \qquad (1)$$

このロジスティック関数で描かれた曲線は項目特性曲線（item characteristic curve）と呼ばれ，各項目に対する反応確率（正解確率）が能力推計値（得点）との関係において常に増関数になっていることを示すものである．個々人の全項目に対する反応パターンと式（1）を用いて，スケールスコア（古典的テスト理論の総得点）の推計を行う．そのプロセスでは，個々の項目及び全項目に対する反応パターンを測定単位として取り扱い，各反応パターンを最も頻繁に創出するスケールスコアを式（1）のロジスティック関数を満足させる 2 つの項目パラメタ（a_i 及び b_i）と同時に推計する．ここで P_i は項目 i へ正解（1 を正解，0 を不正解と仮定）する確率，θ_j は被験者 j の能力推計値，e は自然対数である．この項目反応モデルは 2 項モデルであるため，式（1）は能力推計値が θ_j である被験者 *j* が項目 *i* に 1 と反応する確率を

示しており，その確率は被験者の特性値（θ_j）と項目の識別力パラメタ（a_i）及び難易度パラメタ（b_i）で構成されるロジスティック関数で説明できるとするものである．

本研究で用いたデータは，各項目が 4 選択肢で構成されるライカート型の項目であるため，(1)の 2PLM を拡張した Samejima(1969)の GRM（Graded Response Model）を用いて分析した．多項型データを分析する GRM では，特性値 θ を持つ個人 j が，項目 i の選択肢 k か k 以上の選択肢に回答する確率は，式（2）で示される Boundary characteristic function （境界値特性関数）で規定される．

$$P^*_{ik}(\theta_j) = \frac{1}{1+e^{-[a_i(\theta_j - b_{ik})]}} \qquad (2)$$

$$P_{ik}(\theta_j) = P^*_{ik}(\theta_j) - P^*_{i(k+1)}(\theta_j) \qquad (3)$$

項目 i の最も低い選択肢かそれ以上の選択肢に回答する確率は 1.0 であることから，式 (3) で各選択肢に回答する確率を推計する．ここで，$P^*_{ik}(\theta_j)$ は，特性値 θ_j を持つ個人が，項目 i において，選択肢 k 或いは k 以上の選択肢に回答する確率とする．また，a_i は，項目 i に共通の識別力パラメタ，b_{ik} は項目 i の選択肢 k の難易度パラメタとする．

２．方法

本論では，澁谷・渡部・吉村・小久保（2014）において分析に用いられた 2 つの社会調査のデータ（データ 1 及びデータ 2）と同一データが用いられたが，本論で用いられた分析方法は項目反応理論であった．データ 1 は，2012 年に青森市在住の高齢者を対象に行われた社会調査 1 の一部で，データ 2 は，社会調査 1 と同一の対象者を母集団として 2013 年に行われた社会調査 2 の一部である．この 2 つの社会調査の全回答者の内，65 歳以上であった 204 名（調査 1）及び 278 名（調査 2）を分析対象とした．両調査における回答者は，青森地域の町内会などで活発に活動している高齢者で，高齢者の認知レベルを測定する 10 項目の尺度で認知的に問題がないと推定された個人であった．

調査 1 及び調査 2 の両調査で幸福感尺度（5 項目；表 1）と生活の質尺度（18 項目；表 2）が共通の尺度として用いられた．幸福感尺度項目は，表 1 に示されるように，調査 1 と調査 2 で異なるワーディングが用いられた．調査 1 においては全の項目が否定的な表現であったのに対して，調査 2 では 3 項目で肯定的な表現が用いられていた．両尺度の全ての項目は，4 肢選択のライカート型項目で，選択肢 0 は最も否定的な選択肢で，選択肢 3 は最も肯定的な選択肢であった．

第 1 報では，古典的テスト理論を用いた分析として，①信頼性係数（クロンバックの α）の比較，②因子分析を用いた項目の分散寄与率などの比較，③逆転項目を肯定的項目に合わせて逆転して算出した尺度の総合得点の分布比較，④総合得点と生活の質の下位尺度との相関マトリックスの評価などを報告した．本論では，項目反応理論を用いた分析として，情報関数，標準誤差及びカテゴリー反応関数を報告し，第 1 報での古典的テスト理論に基づいた分析と比較する．

本研究で用いた社会調査データの収集においては，身体的危険や人権を著しく侵害する危険のある実験や調査は行われなかったが，心理的外傷などを防ぎ，調査対象者の人権を尊重するため，個人情報保護法を遵守すると共に，全ての調査対象者に調査研究内容を説明し，調査への協力は任意であること及び何時いかなる理由においても，何ら不利益をこ

Table 1 Happiness Item wording for survey 1 and 2

調査　1	調査　2
Q11A：人生を全体的に評価すると，自分は**恵まれていない**と感じる	Q9A：人生を全体的に評価すると，自分は**恵まれている**と感じる
Q11J：自分が不幸であると感じることがある	Q9J　：自分が不幸であると感じることがある
Q11B：普段，幸せであると**感じることは少ない**	Q9M　：普段，幸せであると**感じることが多い**
Q11Y：自分は「不幸な運命に生まれた人間だ」と感じることがある	Q9T　：自分は「不幸な運命に生まれた人間だ」と感じることがある
Q11C：自分は，「生きていてよかった」と**感じることは少ない**	Q9V：自分は，「生きていてよかった」と**感じることが多い**

うむることなしに，自由に研究への協力を断ることが出来ることを説明した．また，上記の社会調査票の内容及び調査プロセスについては，青森大学社会学部の社会調査倫理規定委員会の承認を得た．

3. 結果

3. 1　回答者の属性

回答者の性別，平均年齢，教育レベルなどの属性は，表 2 に示した．調査 2 における男性の割合が 43. 9%（278 人中 122 人）で，調査 1（204 人中64 人；31. 4%）と比較して高かった．既婚者の割合は，調査 2（75. 5%）において調査 1（65. 7%）と比較して若干高かった．既婚者の割合は，両調査において男性のほうが明らかに高かった．調査 1 では，男性の既婚者が 85. 9%（55 名），調査 2 においては，91. 0%（111 名）であった．教育レベルは，調査 1 で 23. 5%が短大卒以上で，調査2 と比較して若干高かった．両調査において，女性の学歴が男性と比較して高い傾向が見られた（調査 1 では，48 名の短大卒以上の学歴の回答者の内，30 名が女性；調査 2 においては，49 人の短大卒以上の内 31 名が女性）．二つの調査間で回答者の属性には，若干の差はあるもの，その他のデモグラフィックな特性には大きな差は見られなかった．

Table 2 Demographic characteristics for respondents

回答者の特性	調査 1（n=204）	調査 2（n=278）
平均年齢	73. 03 (σ=6. 01)	72. 67 (σ=5. 94)
男性の割合	64　（31. 4%）	122　(43. 9%)
既婚者の割合	134　（65. 7%）	210　（75. 5%）
子供いる割合	191　（93. 6%）	259　（93. 2%）
教育レベル	中卒:52（25. 5%） 高卒:94（46. 1%） 短大：48 (23. 5%) 卒以上	中卒:66（23. 7%） 高卒：160 (57. 6%) 短大：49 (17. 6%) 卒以上

本研究における回答者は，ランダムにサンプルされてはいなかったが，同一母集団からサンプルされている事，回答者の属性などが大きく異ならないこと，10 項目の認知テストで回答者の認知能力が確認されていることなどから，項目のワーディングの影響を評価する比較分析が可能であると結論された．

3. 2 幸福感尺度の比較

3. 2. 1 情報関数及び信頼性

古典的テスト理論を用いた分析として，調査1と2の幸福感尺度の信頼性比較を可能にするため，幸福感が高い場合に得点が高くなるように尺度の方向性を整え（調査 1 においては全 5 項目を逆転し，調査 2 においては 2 項目を逆転し），クロンバックの **α** を比較したが，両調査で顕著な相違は見られなかった（$\boldsymbol{\alpha}_{\text{調査 1}}$=. 781，　$\boldsymbol{\alpha}_{\text{調査 2}}$=. 784）．項目内容が抽象的であること，更に項目数が比較的少ない尺度であることを考慮すると，信頼性係数は十分に高いレベルであると評価できる．澁谷・渡部・吉

村・小久保　（2014）は，古典的テスト理論に基づいた分析では，尺度を構成する項目の一部の項目の項目反応を逆転させることは，尺度の信頼性に大きく影響を及ぼしていなかったと結論している．

しかし，項目反応理論を応用し，情報関数と標準誤差を用いた分析では，二つの調査における相違が見られた．情報関数は，尺度構成に用いられた項目群に対する回答者の反応パターンに基づいて抽出された情報量を特性値θ（シータ）と呼ばれる，古典的テスト理論における総得点にあたる数値の関数で表現したもので，特性値毎に抽出される情報量を評価したものである．古典的テスト理論では，測定値の標準誤差は尺度全体で同一の標準誤差とされているが，項目反応理論においては，特性値θ毎に抽出された情報量が異なるため，特性値θ毎に標準誤差が推計される．この点は，項目反応理論が古典的テスト理論と比較して優れているとされる点で，クロンバックのαを用いた古典的テスト理論分析では得ることができない情報である．選択肢の各カテゴリーの抽出情報量は，カテゴリー情報関数（式4）に示されるように I_{ik}（θ）で表現される．式4における P_{ik}（θ）は能力がθである回答者が項目 i のカテゴリー k に反応する確率である．項目全体で抽出される情報量は項目情報関数（式5）で示され，用いられた項目のカテゴリー情報関数の総和となる．

$$I_{ik}(\theta) = \frac{\partial^2 \log P_{ik}(\theta)}{\partial(\theta)^2} \qquad (4)$$

$$I_i(\theta) = \sum_{k=1}^{k} I_{ik}(\theta) P_{ik}(\theta) \qquad (5)$$

尺度全体で抽出できる情報量は，テスト情報関数で表され，項目情報関数の総和となる．

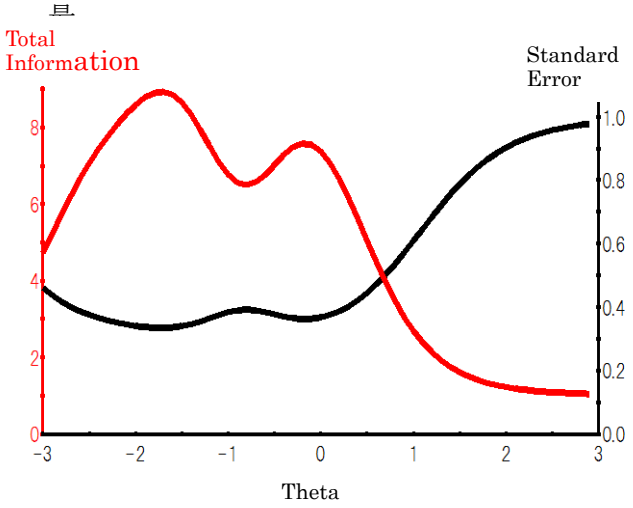

Figure 1 Total information function and corresponding standard errors in survey 1

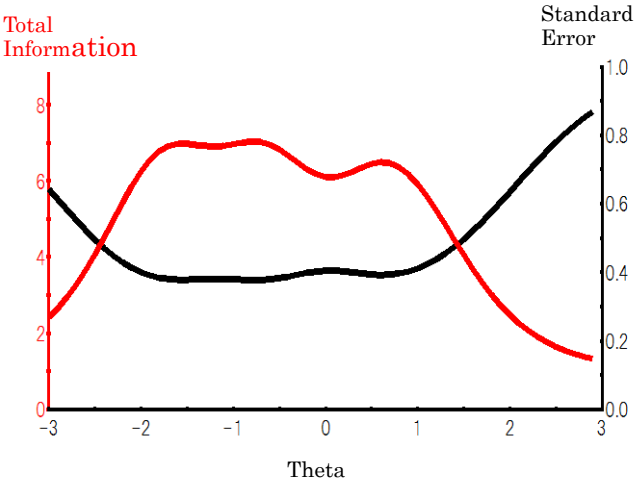

Figure 2 Total information function and corresponding standard errors in survey 2

調査1と調査2における Grated Response Model (GRM) を用いたテスト情報関数及び測定値の標準誤差を図―1及び2に示した．幸福感尺度で抽出された情報量は，調査1尺度では，x 軸の特性値θ尺度（幸福感尺度）で低値である - 3.0（偏差値換算で 20 点）で既に 5.0 と高く，θ値の上昇とともに 9.0 程度まで増加し，θ値で 0.0（偏差値換算で 50 点）から急激に低下し，θ値で＋1.0 では 2.0，θ値が+2.0

では情報量は 1. 0 程度まで低下した（図 1）. 一方，調査 2 尺度の抽出情報量は, θ 尺度の最低値である - 3. 0 では, 2. 5 程度と低かったが, θ 値で約-2. 0 では約 6. 0 まで上昇し, θ 値で約＋1. 0 程度までは 6. 0 以上の抽出情報量を示した. しかし，抽出情報量は, θ 値で＋1. 5 では約 4. 0 まで低下し，それ以後は急激に減少し, θ 値+3. 0 では 2. 0 以下にまで落ちた（図 2）. 調査 1 の幸福感尺度は，特性値 θ が低いレベルで情報抽出量が高かったのに対して，調査 2

の尺度では幸福感が高いレベルでの情報抽出量が高く，両尺度で明確に異なる情報抽出特性を示した. 更に，調査 1 の尺度は，抽出情報量の最高値が約 9 であり，最高値が約 7 であった調査 1 と比較して高かった.

調査 1 の幸福感得点の平均は，古典的テスト理論の分析では 19. 75 (σ = 4. 00)で，調査 2 の平均（16. 21 ; σ = 2. 96）と比較して有意(p<. 01) に高かった（澁谷・渡部・吉村・小久保（2014））. 項目反応理論の特性値 θ は，調査 1 の平均を 0. 0，標準偏差を 1. 0 としたスケールを用いて推計され，調査 2 の平均は 0. 60（σ = . 98）と推計され，平均値の比較では，古典的テスト理論の分析と同様の傾向であった. 項目反応理論を用いた分析には, Samejima（1969）の Graded Response Model が用いられ，パラメタの推計結果は表 3 及び 4 に示した. また，典型的なカテゴリー特性関数を示した項目 11b を 図 3 及び 4 に示した. 図 3 及び 4 では，情報関数は破線で示されているが，5 項目全体の情報関数を

Table 3 Estimated item parameters for survey 1

Item	Label	a_i	b_1	b_2	b_3
1	q11a	1. 65	-2. 41	-1. 47	-0. 04
2	q11b	2. 83	-2. 58	-1. 54	0. 23
3	q11c	2. 41	-2. 40	-1. 76	-0. 26
4	q11j	3. 11	-1. 91	-1. 48	-0. 25
5	q11y	1. 32	-3. 35	-2. 33	-0. 63

Table 4 Estimated item parameters for survey 2

Item	Label	a_i	b_1	b_2	b_3
1	q11a	1. 54	-2. 17	-0. 96	1, 18
2	q11b	2. 34	-1. 67	-0. 60	1. 03
3	q11c	2. 68	-1. 84	-0. 62	0. 57
4	q11j	2. 06	-1. 59	-0. 91	0. 68
5	q11y	1. 66	-1. 77	-0. 82	0. 53

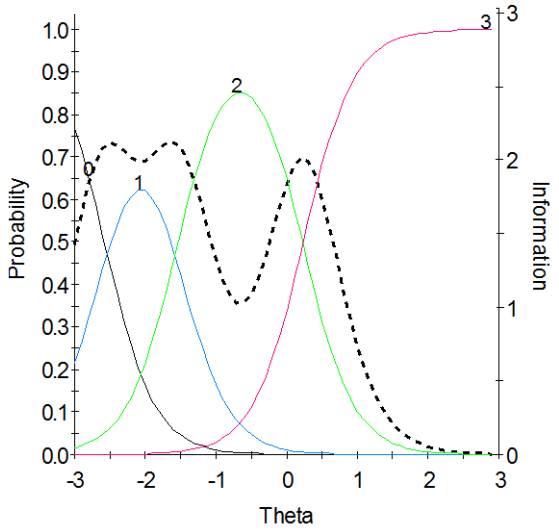

Figure3. Category characteristic functions and corresponding standard errors for q11b in survey 1

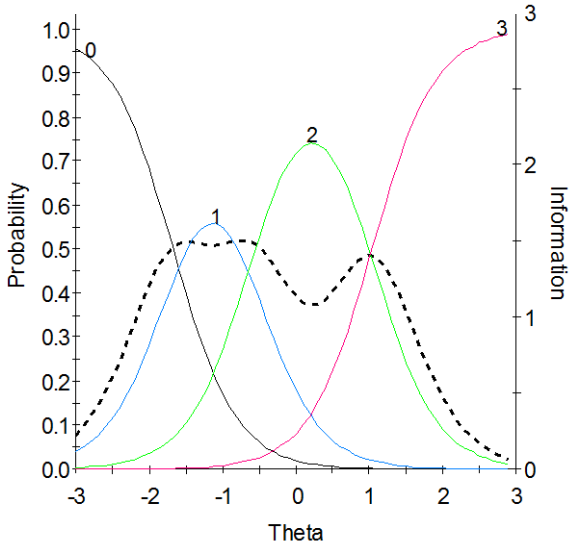

Figure4. Category characteristic functions and corresponding standard errors for q11b in survey 2

加算した図2及び3のテスト情報関数と同様の傾向を示している．調査 1（図 3）においては，特性値 θ で-2. 5 から-1. 5 周辺の抽出情報量は，2. 0 付近であるが，-. 75 周辺では 1. 0 まで低下しているが，特性値 θ が. 5 周辺では 2. 0 付近まで増加し，特性値 θ が+. 5 周辺では抽出情報量はほぼ 0 まで低下している．一方，調査 2（図 4）では，特性値 θ で-1. 75 から+1. 5 の範囲では，抽出情報量は 1. 0 以上を維持しており，より広い特性値 θ の範囲で情報が抽出されていた．

特性値が高い回答者と低い回答者を識別する項目識別力は，a_i パラメタ（item discrimination parameter）で示され，通常は 0～3 程度の範囲である．識別力は，高いほど測定される特性値を識別する能力が高い．調査 1 では 5 項目中 3 項目の項目識別力が調査 2 と比較して高かったが，全体として大きな相違はなかった（表 3, 表 4）．

特性値 θ を持つ個人 j が， 項目 i の選択肢 k か k 以上の選択肢に回答する確率は， 境界値特性関数で規定され，本調査の項目における選択肢数は 4 であるため，各選択肢の境界値特性関数の交差する点の特性値 θ は，b_1 から b_3 で示されている（表 3, 表 4）．境界特性値である b_i は，回答者が隣接する選択肢を選択する確率が同等となる特性値 θ のことで，例えば，表 1 の b_1 は，-2. 41 であることから，特性値 θ が-2. 41 未満の回答者は選択肢 1 を選択する確率が高く，特性値 θ が -2. 41 より高い回答者は選択肢 2 を選択する確率が高くなることを示す．

調査 1 の項目 q11b の境界特性値は，低い特性値 θ に対応する選択肢 0 と 1 の識別指標である b_1 は -2. 58 であるのに対し，調査 2 の b_1 は -1. 67 であり，かなり高かった．この傾向は他の項目・選択肢でも同様の傾向を示した．調査 1 では，特性値 θ が低い部分で正確な測定が行われ，調査 2 では特性値 θ が比較的高い部分での正確な測定が実施されていたといえる．この結果は，各選択肢に回答する確率を示すカテゴリー反応曲線（図 3・4）の形状からも支持されるもので，調査 1 の項目 q11b では抽出情報量が高い特性値域において，明確に選択肢 0 か 1 に対応する特性値 θ は，-2. 75 から-1. 5 付近に集中しており，オーバーラップする部分が多かった（図 3）．しかし，調査 2 の項目 q11b の境界特性値では，抽出情報量が 1. 0 以上の特性値レンジは，約-2. 5 から+1. 25 と幅広く，4 本のカテゴリー反応曲線は，特性値-3 から+3 までほぼ均等に分布し，特性値のレベルによって回答者が選択する選択肢が比較的判断しやすかった（図 4）．

4. 結論

本研究において分析された幸福感尺度は，調査 1 と調査 2 でワーディングが異なる項目が 3 項目あり，調査 1 では全 5 項目が否定的表現を用いた項目であったが，調査 2 では 2 項目が否定的項目であった．調査 1 では，全 5 項目を逆転し，調査 2 では否定的項目 2 項目を逆転させ幸福感の総得点を計算した．これらの 2 調査間の相違が尺度に及ぼす影響は，古典的テスト理論による分析では，下記の 2 項が報告されている（澁谷・渡部・吉村・小久保 ; 2014）．

（1） 幸福感得点の平均値は，調査 1 （平均値=19. 75 ; σ=4. 00）の方が調査 2（ では 平均値=16. 21 ; σ= 2. 96）と比較して有意に 高かった (p<. 01).

（2） 調査 1 と調査 2 において，クロンバックの α に基づいて行われた信頼性の比較では，顕著な相違は見られなかった（$\boldsymbol{\alpha}_{\text{調査 1}}$=. 781, $\boldsymbol{\alpha}_{\text{調査 2}}$=. 784）．

本論では項目反応理論を用いて尺度分析を行った結果，下記の 4 項が明らかとなった．

（1） 情報関数と測定値の標準誤差を用い

て評価した信頼性によると，調査１の信頼性は調査 2 と比較して高かったが，抽出情報量が高い特性値 θ の範囲は特性値 θ が低い部分であった．

（２） 調査 2 の尺度全体としての抽出情報量は，調査 1 と比較して特性値 θ の広い部分をカバーしており，特性値 θ が+. 5以上の範囲では，調査2の抽出情報量は調査１と比較して明確に高かった．

（３） 特性値 θ の平均値は，調査１（平均値=0. 00，σ = 1. 00）の方が調査２（平均値=0. 60，σ = . 98）と比較して明確に高かった．

（４） 調査 1 のカテゴリー反応関数（図 3 及び 4）は，特性値 θ の低い部分に分布し，調査2では特性値θの広い範囲にわたる分布を示した．

O’Muircheartaigh, Krosnick ＆ Helick（2000）は，肯定的な項目で構成される尺度の中に否定的な項目が混在すると，尺度の信頼性が低下する傾向があることを報告している．O’Muircheartaigh, Krosnick ＆ Helick（2000）は，澁谷・渡部・吉村・小久保（2014）の研究と同様に，信頼性の評価にクロンバックの α を用いたが，澁谷らの研究では，二つの尺度間にはクロンバックの α による信頼性の大きな相違は見られなかった．しかし，本研究と全く同様の２尺度を用いた澁谷らの研究で用いた尺度の項目数は，O’Muircheartaigh, Krosnick ＆ Helick（2000）の用いた項目数より少ないこと，さらに，澁谷らの研究では項目内容の抽象性が高いことが信頼性の評価に影響している可能性がある．

本研究では，否定的項目で構成されている尺度（調査 1）の情報関数は，特性値 θ の低い部分に限定されていたが，調査2と比較して高かった．しかし，肯定的な尺度（調査 2）は，情報関数のピークは低かったものの，特性値 θ の幅広い領域に比較的高い情報関数の分布が見られた．この結果は，否定的項目の方が情報抽出力は高いが，その情報抽出力は限定された特性値 θ の領域において起こっていることを示すもので，この特徴を有効に用いることで尺度を改善することができる可能性を示唆するものである．

Chen, Rendina-Gobioff & Dedrick（2007）は，否定的項目と肯定的項目を混在させると尺度の構成概念妥当性が低下する可能性があると報告している．澁谷・渡部・吉村・小久保（2014）の研究では，項目のワーディングの相違は，平均値の有意な相違を起こす可能性はあるが，それが直ちに幸福感尺度の構成概念妥当性を損なうとする論拠を提供するまでは至らないとしている．一般的な社会調査においては，30 項目にも及ぶ多数の項目で構成される尺度が使用されることは少なく，何らかの尺度が用いられるとしても，多くの場合，用いられる尺度は 5〜10 項目程度で構成されている．本研究では，否定的な項目で測定した場合には，カテゴリー反応関数が低い特性値 θ に集まる傾向があるが抽出情報量は高く，肯定的項目で測定した場合には，抽出情報量は比較的低いものの，カテゴリー反応関数は，特性値 θ の広い範囲に分布することが明らかとなった．さらに，カテゴリー反応関数の分布などから，項目のワーディングを変化させることが，尺度の妥当性が直ちに大きな問題となるとは考えにくい. Liez (2010) は，項目の肯定性･否定性自体が因子負荷量の増減の原因となるため，尺度を構成する因子構造に重要な影響を及ぼすと報告しているが，項目反応理論を用いた分析結果は，項目内容が抽象的で項目数が少ない尺度における否定的な項目は，

抽出情報量の観点から尺度の改善に寄与する可能性があることを示唆するものであった.

本研究の結果から，肯定的項目と否定的項目を同一尺度に混在させることが尺度の妥当性及び信頼性に及ぼす影響は，古典的テスト理論だけでは明確とならない部分があり，項目反応理論に基づく分析が有効であった．否定的項目は情報抽出量が高い傾向にあったが，情報抽出量は特性値 θ が平均値より低い部分に集中していたことから，肯定的項目を混在させ尺度全体で高い情報抽出量を確保することが望ましいと考えられる．肯定的項目と否定的項目を混在させて尺度を構成する場合，項目反応理論を用いた分析で妥当性が確保できていることを確認する必要がある．また，信頼性についても情報関数などを用いて，測定目的に必要な情報量が抽出されていることを確認する必要がある.

本研究の分析結果は幸福感に関するものであり，測定される心理学的・社会学的特性に一般化できるか否かについては，研究を蓄積していく必要がある.

また，実際に行われている社会調査では，項目数が少ない尺度を用いることが一般的である．社会調査の実情に鑑み，少ない項目で構築される尺度における項目ワーディングの影響や，通常の日本語表現で否定的な言い回しをする表現を調査項目にする場合，その表現を肯定的な表現に変換すべきか，などは更に研究が必要である.

＊調査 1 は，平成 23 年度公益財団法人日工組社会安全財団の助成を受けて行われた研究の一部である．また，調査 2 は，科学研究費補助金（課題番号 : 23530825）の助成を受けて行われた研究の一部である.

References

Anastasi, A. & Urbina, S. (1997). *Psychological testing*, Upper Saddle River,

Biemer, P. , P. & Lyberg, L. E (2003). *Introduction to Survey quality*, Hoboken, Wiley

Chen, Y. ,Rendina-Gobioff, G., Dedrick,R. F. ,(2007). Detecting effects of positively worded items on a self-concept scale for third and sixth grade elementary students, *Paper presented at the annual meeting of the Florida Educational Research Association*, Tampa, FL. Nov. 14-17.

Dillman, D. A. (1978). *Mail and telephone surveys: the total design method*, Wiley & Sons.

Liez, P. P. (2010). Factor analysis and negatively worded items, *International journal of market research*, 52(2), 249-272.

Nunnally, J. C. & Bernstein,I. H. (1994). *Psychometric theory (3rd ed.)*, New York, NY: McGraw-Hill.

O'Muircheartaigh, C. , Krosnick, J., & Helick, A. (2000). Middle alternatives, acquiescence, and the quality of questionnaires, *unpublished manuscript, available at harrisschool. uchicago. edu/about/publications/working-papers/pdf/wp _01_3. pdf.*

澁谷泰秀・渡部諭・吉村治正・小久保温, (2014). 肯定的項目と否定的項目が社会調査データの分析に及ぼす影響：古典的テスト理論を用いた分析, 『青森大学付属総合研究所紀要』16(1), 1-13.

Spector, P. E., (1992). *Summated rating scale construction; introduction*, Sage University paper series. Quantitative applications in the social sciences No. 82, Newbury Park, CA: SAGE publications, Inc.

Tversky, A. , and Kahneman, D. (1981). The framing of decisions and the psychology of choice, *Science*, 211, 454-458.

von Neumann, J., & Morgenstern, O. (1947). *Theory and games and economic behavior (2nd ed.)*, Princeton, NJ: Princeton University Press.

Evaluating the influence of mixing positively and negatively expressed items in a scale: Social survey data analysis utilizing item response theory

Hirohide SHIBUTANI[1]　Satoshi WATANABE[2]
HarumasaYOSHIMURA[3]　Atushi KOKUBO[4]

[1] Faculty of Sociology,　Aomori University
[2] Research and Education Center for Comprehensive Science,　Akita Prefectural University
[3] Faculty of Sociology,　University of Nara
[4] Faculty of Software and Information Technology,　Aomori University

要旨

本研究の目的は，社会調査における項目のワーディングの相違が尺度の妥当性及び信頼性に及ぼす影響を評価することである. 2011 年と 2013 年に行われた調査 1 及び調査 2 において用いられた，二つの幸福感尺度（5 項目で構成され，5 項目全てが否定的表現で構築されていた調査 1 と 3 項目が肯定的で残り 2 項目が否定的な項目で構成された尺度が用いられた調査 2）を項目反応理論（Graded Response Model）で比較した．全ての否定的項目は，幸福感尺度の方向性を整えるために逆転して得点を算出した．否定的項目のみで構成された尺度の特性値 θ の平均値は，項目表現が混在した尺度と比較して高かった．否定的項目のみで構成された尺度は，抽出情報量は高かったが，抽出情報量が高いスケールスコアの範囲は 0 以下の低い部分に集中していた．項目表現が混在した尺度は，抽出情報量は比較的低かったが，特性値 θ の広い範囲で情報が抽出され，それに対応してカテゴリー反応関数は特性値 θ の広い範囲にわたる分布を示した．結論として，肯定的項目と否定的項目を混在させ尺度全体で高い情報抽出量を確保することが望ましいことが示唆された．

キーワード：項目ワーディング，肯定的項目，否定的項目，項目反応理論，段階反応モデル