

特異項目機能検出方法の比較

— BILOG-MG と SIBTEST を用いた検討 —

熊谷 龍一¹⁾ 脇田 貴文¹⁾

問題と目的

特異項目機能 (differential item functioning: DIF) は、「能力が同一であるにもかかわらず、所属する集団の違いから、正答率 (反応確率) に差が生じる項目」と定義される (渡辺・野口, 1999)。例えば、性別・地域・民族などの違いにより、片方の集団に有利 (もしくは不利) な項目となる時、その項目について DIF が生じているとされる。そのような項目を含めたままテスト・尺度を分析した場合は当然のことながら、片方の集団に有利あるいは不利な結果を導くことになる。そのため、この DIF が生じている項目を検出することは、テスト・尺度を開発する際に重要である。

DIF に関する研究は、主に米国で活発に行われており、DIF を検出するための方法も様々なものが提案されている。

DIF 検出の方法は「パラメトリックな方法」と、「ノンパラメトリックな方法」とに分類することができる。「パラメトリックな方法」とは、項目反応理論 (item response theory: IRT) に基づき、項目パラメタを推定し、それを用いて DIF の検出を行うものである。具体的には、Lord の χ 自乗法 (1980)、尤度比を用いた方法 (Thissen, Steinberg, & Wainer, 1993)、そして項目パラメタ推定プログラムである BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) に用意されている DIF コマンドを使用する方法などがある。「ノンパラメトリックな方法」は、パラメタの推定という手続きを行うこと無しに DIF を検出する方法で、SIBTEST (Shealy & Stout, 1993) による方法、Mantel-Haenszel 統計量を用いる方法などがある。

日本では、これまで項目バイアスや DIF への関心はそれほど高くなく、実際の応用研究の例も多くは無い。そのため、実際に DIF 分析をする際に、どのようなデー

タに対してどの DIF 検出の方法を用いればよいのか、といった実際的な適用に関する知見の集積がまだ十分とは言えない状況である。そこで本研究では、実際に DIF 分析を行う場面で用いられるであろう DIF 検出の方法についての比較検討を行い、今後の DIF 分析に対して何らかの示唆を与えるような知見を提示することを目的とする。

本研究では、「パラメトリックな方法」として BILOG-MG での DIF コマンドを、「ノンパラメトリックな方法」として SIBTEST を取り上げ、比較検討を行う。これら 2 つの方法は、それぞれ専用のプログラムパッケージが用意されており、容易に実行可能なものである。

比較検討では、はじめにシミュレーションデータを用いる。シミュレーションデータ発生の際に、全項目に対する DIF が生じている項目の割合、被験者母集団の能力分布のずれなどの要因を変化させることで、これらの要因が各 DIF 検出にどの程度の影響を与えるのかを検討する。

次に、実際データにおける DIF 検出方法の比較を行う。実際のデータには、通常シミュレーションデータでは想定されない様々な要因 (誤差) が混入している。そのような要因が、実際の DIF 検出の方法にどのような影響を与えているのかについて検討する。実際データには、芝 (1978) による語彙理解尺度を、大学生集団と専門学生集団に実施したもの (熊谷, 2003) を用いた。

研究 1 シミュレーションデータによる比較

目的

BILOG-MG と SIBTEST との比較を、シミュレーションデータを用いて行う。シミュレーションデータの場合、真のパラメタを予め設定することが可能であり、2 つの集団間で困難度パラメタに差を与えることで、DIF が生じている項目を設定することができる。

本研究では、

1. 全体項目 20 項目に含まれる DIF 項目の数
2. 各 DIF 項目における真の困難度パラメタの差

1) 名古屋大学大学院教育発達科学研究科博士課程 (後期課程)

3. 2群間における潜在特性値の母集団分布の違いに焦点をあてる。

方法

本研究では全部で20項目のテストを想定し、DIFを生じさせた項目数に関して3通り（2項目、6項目、10項目）、困難度パラメタの差に関して3通り（0.5、0.7、1.0）、母集団分布の差に関して3通り（0.0、1.0、2.0）の計27個（3×3×3）のデータセットを用いた。シミュレーションデータの生成には RESGEN 4（Muraki, 2001）を用いた。

各データセットの生成の手続きは以下の通りである。始めに共通項目として20項目分の項目パラメタを、識別力パラメタに関しては平均0.00、標準偏差0.15の対数正規分布、困難度パラメタに関しては標準正規分布からランダムに発生させた。この手続きにより得られたパラメタを、基準の項目パラメタ値とした（Table 1）。そして、DIF 項目についてはこの基準となる困難度パラメタに下位集団間で差が生じるように真の値を設定した。ここでは、DIF 項目が2項目の場合は、項目19、項目20、6項目の場合は項目15から項目20、10項目の場合は項目11から項目20を DIF が存在する項目とし、基準となる困難度パラメタに0.5、0.7、または1.0を加えた。例として DIF を生じさせた項目数6、困難度パラメタの差1.0の場合についてのパラメタ値を、Table 2 に示した。

これらのデータセットに対して BILOG-MG と SIBTEST による DIF 検出の分析を行った。

BILOG-MG による DIF 検出の方法 BILOG-MG では DIF 検出を行う際に、DIF コマンドを使用するが、その場合、均一 DIF のみを扱っている。したがって、項目パラメタを推定する際には識別力パラメタに関して

は両群で同一の値に固定され、困難度パラメタは2群で個別に推定される。その推定後に、全項目の困難度パラメタの平均を各群で算出し、その差を解消するように等化調整を行う。そして、等化調整後の各項目困難度パラメタ値が2群間で一定程度以上の差が見られる場合に DIF が生じているとする。

SIBTEST による DIF 検出の方法 以下の手続きにより、SIBTEST による DIF 検出を実施した。

1. 各項目について、その項目以外の項目を DIF が生

Table 1 基準項目パラメタ

項目番号	識別力	困難度
1	0.847	-1.195
2	0.978	0.739
3	1.057	0.120
4	0.787	2.413
5	1.034	-0.214
6	1.008	-1.759
7	1.082	1.172
8	0.912	-0.468
9	0.994	0.581
10	1.026	-0.611
11	1.101	-1.326
12	0.854	0.372
13	1.226	0.365
14	0.821	-1.417
15	0.931	-0.186
16	1.160	1.267
17	0.778	-2.269
18	1.090	1.147
19	0.988	-0.381
20	0.768	-1.460

Table 2 シミュレーションデータの項目パラメタ例

項目番号	識別力	困難度	項目番号	識別力	Group 1 の困難度	Group 2 の困難度
1	0.847	-1.195	11	1.101	-1.326	-0.326
2	0.978	0.739	12	0.854	0.372	1.372
3	1.057	0.120	13	1.226	0.365	1.365
4	0.787	2.413	14	0.821	-1.417	-0.417
5	1.034	-0.214	15	0.931	-0.186	0.814
6	1.008	-1.759	16	1.160	1.267	2.267
7	1.082	1.172	17	0.778	-2.269	-1.269
8	0.912	-0.468	18	1.090	1.147	2.147
9	0.994	0.581	19	0.988	-0.381	0.619
10	1.026	-0.611	20	0.768	-1.460	-0.460

じていない項目 (matching item) として、 β 推定量と呼ばれる指標を算出する (このときの当該項目のことを suspect item と呼ぶ)。

2. 手続き1で β 推定量の絶対値が0.1を超えている項目を DIF が生じている項目と見なし、次の手続きから削除する。残った項目に対して再び手続き1を行い、DIF が生じている項目がなくなるまでこの手続きを繰り返す。
3. 手続き 1, 2 を通して残った項目を matching item 群とし、手続き 2 で削除された項目に対して個別に β 推定量を算出する。この β 推定量の絶対値が0.1を超えた項目が、最終的に DIF が生じていると判断する。

なお、SIBTEST の計算には、DIF-PACK (Stout, 1999) を用いた。このプログラムでは、 β 推定量が0であるという検定仮説に対する検定が実施される。しかし、今回のシミュレーションデータのように受験者データ数が多い場合には、p 値が極端に小さくなりほぼ全ての項目が棄却され、DIF が生じている項目と判定されるため、上述の手続きのように β 推定量の値を解釈することで DIF の検討を行った。また、 β 推定量の絶対値0.1という DIF 判定の基準については、SIBTEST と Mantel-Haenszel 統計量の比較を行った Roussos & Stout (1996) によるものを採用した。

結果と考察

BILOG-MG BILOG-MG の DIF コマンドを用いる場合、どの程度の差をもって DIF とみなすかという点については現在のところ一致した見解が得られていない。Table 3 に潜在特性値の母集団分布の差が1.0、かつ困難度パラメタの差が1.0の場合における、2 群の困難度パラメタ推定値の差を示す。

DIF が生じている項目が 2 項目の場合、項目 1 から項目18までは、2 群の差の絶対値が0.011から0.163の間であり、項目19と項目20に関しては0.892, 0.809と大きな差が見られた。この結果から、これら 2 項目について DIF が生じていると判断することができる。また、DIF が生じている項目と DIF が生じていない項目では、データを生成する際に困難度パラメタを1.0ずらした。従って、項目19, 項目20において0.8程度のずれが確認されたことで、適当な推定がなされていると考えられる。

同様に、DIF が生じている項目が 6 項目の場合、項目 1 から項目14までは、おおよそ0.3程度であり、項目 15から項目20までは、0.7程度の差がみられた。DIF が生じている項目が 2 項目の場合ほど極端ではないが、DIF が生じている項目とそうではない項目を識別する

ことが可能である。しかし、DIF が生じている項目に関して 2 群間で1.0の差を設定したにもかかわらず、0.7程度の差しか見られなかった。

DIF が生じている項目が10項目の場合は、DIF が生じていない項目 1 から 項目10と DIF が生じている項目 11から項目20では正負の符号は異なるものの、絶対値を取るとどちらもほぼ0.5付近を示しており両者の区別は困難である。

これらの結果から、DIF 項目が 6 項目の場合、10 項目とその割合が増えるにつれて、DIF 項目とそれ以外の項目で差の絶対値に違いが見られなくなり適切な推定が行われていないことが示された。これは、BILOG-MG によるパラメタ推定では SIBTEST における matching item を設定しておらず、DIF が生じている項目も含めてパラメタの推定を行っているためであると考えられる。したがって、DIF 項目の割合が高い場合には適切なパラメタの推定がなされない可能性が高くなるため留意することが必要である。

また、BILOG-MG では困難度パラメタの差をもって DIF の検討を行うが、両群のパラメタの差がどの程度であれば DIF があると判断するかの基準が明確ではな

Table 3 シミュレーションデータにおける BILOG-MG の結果

項目番号	2 項目	6 項目	10項目
1	-0.090	-0.378	-0.508
2	-0.081	-0.309	-0.458
3	-0.148	-0.329	-0.498
4	-0.011	-0.236	-0.512
5	-0.111	-0.255	-0.486
6	-0.156	-0.368	-0.560
7	-0.138	-0.336	-0.543
8	-0.163	-0.250	-0.503
9	-0.077	-0.288	-0.521
10	-0.105	-0.335	-0.502
11	-0.103	-0.263	0.471
12	-0.060	-0.301	0.547
13	-0.109	-0.306	0.496
14	-0.106	-0.265	0.589
15	-0.123	0.672	0.445
16	-0.080	0.740	0.501
17	0.049	0.705	0.525
18	-0.087	0.733	0.542
19	0.892	0.716	0.507
20	0.809	0.652	0.468

網掛けは DIF が生じていると設定した項目

特異項目機能検出方法の比較

い。シミュレーションデータでは、DIF 項目が既知であるため DIF 検出の適切性に関する判断が可能であるが、実データの場合にはそのような判断をすることができない。

潜在特性値の母集団分布の差の大きさに関して DIF 項目の数が同一の場合で比較した場合、その影響はほと

んど見られなかった。例として、困難度パラメタの差が 1.0, DIF が生じている項目が6項目の場合を Table 4 に示した。

Table 4 母集団分布の違いによる BILOG-MG の結果

項目番号	母集団分布の差		
	0.0	1.0	2.0
1	-0.321	-0.378	-0.358
2	-0.276	-0.309	-0.310
3	-0.266	-0.329	-0.312
4	-0.248	-0.236	-0.326
5	-0.280	-0.255	-0.315
6	-0.368	-0.368	-0.147
7	-0.328	-0.336	-0.320
8	-0.312	-0.250	-0.245
9	-0.287	-0.288	-0.347
10	-0.281	-0.335	-0.315
11	-0.333	-0.263	-0.286
12	-0.308	-0.301	-0.266
13	-0.317	-0.306	-0.340
14	-0.267	-0.265	-0.213
15	0.740	0.672	0.647
16	0.718	0.740	0.733
17	0.718	0.705	0.713
18	0.713	0.733	0.700
19	0.712	0.716	0.760
20	0.589	0.652	0.549

SIBTEST SIBTEST による DIF 検出の結果を Table 5 に示した。各セルには DIF が検出された項目番号を示した。

BILOG-MG の時と同様に、母集団分布の差が1.0の場合を見ると、困難度パラメタの差が0.5の場合は、DIF が生じている項目が2項目の場合に1項目、6項目の場合に2項目、また10項目の場合には0項目と適切な検出がなされていない。これは困難度パラメタの0.5という差が、項目反応パターンに影響するほど大きいものでなかったためであると考えられる。このことは、DIF 項目が2項目、6項目の場合には、困難度パラメタの差が0.7, 1.0になるにつれて適切な検出がなされていることから明らかである。

ただし、DIF が生じている項目が10項目（全項目の半分）の場合は、DIF が生じていないと設定した項目について DIF が存在すると検出している。特に、困難度パラメタの差が1.0の場合には、逆の結果を導いている。そこで、DIF が生じていないと仮定した項目（項目1から項目10）を matching item とし、残りの項目を suspect item として分析したところ、すべての項目について DIF が生じているという結果が得られた。つまり、上述した DIF が生じているとした項目が10項目の場合に逆の結果となった原因としては、matching item を決定していく過程において、DIF が生じていると仮定した項目を matching item と判断したためであると考えられる。したがって、SIBTEST による DIF 検出には、どの項目が matching item とされるかが、

Table 5 シミュレーションデータにおける SIBTEST の結果

母集団分布の差	困難度パラメタの差	DIF を生じさせた項目数		
		2	6	10
0.0	0.5	19	15, 19	X
	0.7	19, 20	15, 18, 19, 20	1~10
	1.0	19, 20	15~20	1~3, 5~10
1.0	0.5	19	15, 19	X
	0.7	19, 20	15, 16, 18, 19	2, 3, 7, 12, 13, 15, 19
	1.0	19, 20	15~20	1~3, 5, 7~10
2.0	0.5	19	15, 16, 18	X
	0.7	19	15, 16, 18, 19	11~19
	1.0	19, 20	15, 16, 18~20	Failure

X は DIF がひとつも検出されなかったことを示す。

Failure は matching item が少数になり、分析が行えなかったことを示す。

大きく影響している。

潜在特性値の母集団分布の影響をみると、DIF が検出された項目に多少の違いがみられた。しかし、母集団分布の差が大きくなることで、DIF 検出項目数が増加するといったような傾向は見られず、母集団分布の違いが DIF 検出に与える影響は小さいといえる。

研究 2 実際データによる比較

目 的

ここでは実際データを用いて、BILOG-MG の DIF コマンドによる方法と、SIBTEST による比較を行うことを目的とする。今回の研究で用いられたデータは、芝 (1978) による語彙理解尺度を大学生集団と専門学生集団に実施して得られたものである。このデータに対して、研究 1 で行われた手続きと同様に DIF 分析を行い、両方法の比較検討を行う。

方 法

テスト構成 語彙理解尺度は、小学 1 年生版から大学生版までの 11 の下位尺度から構成されており、また芝・野口 (1982) により等化の手続きが行われている。本研究では、この等化された項目パラメータを用いて、大学生版語彙理解尺度の中から、困難度及びテスト情報量を考慮しながら 50 項目を選択し、一つのテストを構成した。

受験者 大学生集団は愛知県内の大学生および大学院生の 322 名であり、専門学生集団は愛知県内の看護系専

門学校生 243 名であった。

手続き 両集団ともに教室での一斉実施という形式でテストを実施した。

結 果

基礎統計量および尺度の 1 次元性の確認 両集団ごとに、各項目の通過率、標準偏差、項目-尺度間相関を算出した。そこで通過率が .05 以下もしくは .95 以上であった 4 項目については以後の分析から除外した。また、項目-尺度間相関が .10 を下回る項目が 7 項目見られた。通常の分析であれば、このような項目は尺度の 1 次元性から逸脱している項目であると考えられるため、以後の分析から除外されるものである。しかし、DIF の影響により項目-尺度間相関が低められている可能性があり、DIF の検出という本研究の目的からすると、このような項目も含めて分析を行うことが妥当であると判断されたため、以後の分析にもこの 7 項目は含められた。

また、項目相関行列 (相関係数にはテトラコリック相関係数を使用) から算出される固有値の減衰状況から尺度の 1 次元性を確認した。第 1 固有値が 15.46 に対して、第 2 固有値が 1.88 となり、スクリープロットからも、この尺度の 1 次元性がかなり強いことが確認された。

BILOG-MG による DIF 分析 大学生集団を参照集団、専門学生集団を焦点集団とし、シミュレーションデータの時と同様に、BILOG-MG の DIF コマンドを用いて DIF 分析を行った。各項目における、両群の困難度

Table 6 大学生集団と専門学生集団の困難度パラメータの差 (等化調整後)

項目番号	困難度差	項目番号	困難度差	項目番号	困難度差
1	1.345	18	-0.814	37	0.343
3	1.413	19	-1.156	38	1.323
4	1.022	21	0.187	39	0.069
5	-1.197	22	-0.860	40	-1.220
6	0.029	23	0.192	41	1.084
7	-0.161	25	2.407	42	0.543
8	0.408	26	0.616	43	1.083
9	0.071	27	-0.181	44	0.166
10	0.764	28	-1.157	45	-1.390
11	-0.273	29	-1.673	46	-0.019
12	2.075	30	0.820	47	-3.981
13	1.222	31	-0.562	48	-1.221
14	0.065	33	-0.768	49	-1.688
15	-0.519	34	0.257	50	1.754
16	0.177	35	-0.133		
17	0.521	36	-0.984		

Table 7 実際データにおける SIBTEST の結果 (suspect item のみ)

項目番号	Beta estimate	standard error	p-value	項目番号	Beta estimate	standard error	p-value
1	0.434	0.070	0.000	29	-0.053	0.050	0.287
3	0.453	0.068	0.000	30	0.204	0.068	0.002
4	0.320	0.071	0.000	34	0.200	0.051	0.000
5	-0.070	0.063	0.268	36	-0.134	0.051	0.009
6	0.159	0.062	0.010	37	0.199	0.073	0.007
7	0.210	0.070	0.003	38	0.418	0.057	0.000
8	0.211	0.052	0.000	39	0.146	0.073	0.046
12	0.167	0.065	0.011	40	-0.110	0.070	0.115
13	0.316	0.056	0.000	41	0.320	0.062	0.000
14	0.113	0.057	0.047	42	0.204	0.058	0.000
16	0.218	0.071	0.002	43	0.233	0.056	0.000
17	0.143	0.063	0.022	44	0.129	0.073	0.076
21	0.310	0.068	0.000	45	-0.120	0.058	0.037
25	0.226	0.058	0.000	50	0.162	0.049	0.001
28	-0.074	0.032	0.022				

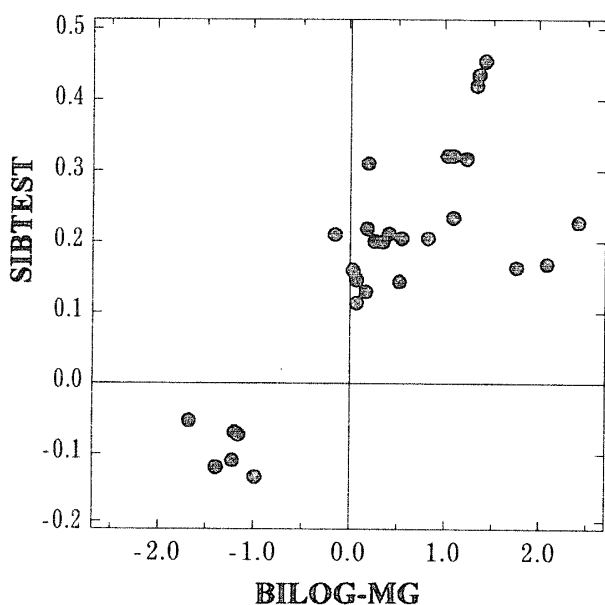


Figure 1 BILOG-MG と SIBTEST の DIF 分析結果の散布図

パラメタの差 (焦点集団-参照集団 : 等化調整済み) を Table 6 に示す。また、この時の大学生集団の潜在特性尺度値の平均は 0.03、専門学生集団の平均は -2.26 であった。

SIBTEST による DIF 分析 シミュレーションデータにおける方法と同様の手続きにより、SIBTEST による DIF 分析を行った。ただし、 β 推定量の DIF 検討基準として ± 0.10 を採用したところ、大半の項目が DIF と判断され、matching item 群が極端に少なくなり、

分析が不可能となった。そのため、DIF 検討基準として ± 0.15 を超えた項目を DIF と判断して分析を行った。その結果、分析対象項の全 46 項目中 16 項目が matching item 群とされ、残りの 30 項目について DIF 検討が行われた。その結果を Table 7 に示す。

両方法の比較 BILOG-MG で得られた困難度パラメタの差と、SIBTEST の β 推定量を比較するために、その散布図を Figure 1 に示す。また、両者の相関は .79 であった。

考察

Figure 1 を見ると、BILOG-MG と SIBTEST による方法はある程度同様な結果を示しているといえる。しかし、Table 6 で示された BILOG-MG の方法の数値を検討していくと、群間で極端な困難度差が生じている項目があることが分かる (たとえば項目 47 など)。また、項目 47 は SIBTEST では matching item と判断されており、DIF が無かったものとなっている。このように、BILOG-MG と SIBTEST においては、DIF が検出される項目にいくつか違いが見られた。その理由として、BILOG-MG には matching item という概念が無く、DIF が生じているかもしれない項目をも含めてパラメタの推定をしてしまうことが挙げられる。研究 I でも示されたように、DIF が生じている項目の全項目に占める割合が大きい場合に、2つの方法間で差が生じると考えられる。

次に、潜在特性尺度値の分布が集団間で異なることの

影響が考えられる。結果で示されたように、大学生集団の尺度値と、専門学生集団の尺度値の差が非常に大きい。これは、テスト構成の手続きにおいて、芝・野口(1982)で推定された項目パラメタをもとに、大学生集団に対して適当な項目を選択したためである。このことから、今回用いられたテストは専門学生に対して非常に難しくなっていたと考えられる。同時に、テスト情報量が受験者の能力尺度値に照らして適切な範囲をカバーしておらず、このことがDIFの検出に関して何らかの影響を与えている可能性がある。

語彙理解尺度の各項目について、なぜDIFが生じているのか、また両集団における違いの原因は何であるのか、という点については項目ごとに精査する必要がある。

総合考察

BILOG-MGのDIFコマンドによる方法と、SIBTESTによる方法とのもっとも大きな違いは、matching itemの存在である。BILOG-MGでの方法では、matching itemという概念が無く、シミュレーションデータの例でも見られるように、まったくDIFが生じていないと設定された項目でも、困難度パラメタに差が生じてしまう。そこで、BILOG-MGにおいても、SIBTESTにおける手続きと同様にmatching itemの検証からはじめるべきだと考えられる。このためには、当該項目以外の項目は両集団で共通項目とし、当該項目のみを別の項目として扱い、パラメタ推定を行う必要がある。

次に、全項目数に対するDIFが生じている項目の割合についてである。DIFが生じている項目が、全項目数の50%を超えるようになると、適切にDIFの検出が行えていないことが分かる。シミュレーションデータでは、matching item群とsuspect item群が逆転するという現象が起きていた。しかし、実データの場合には、実際にどの項目にDIFが生じているのか(もしくは、DIFが生じていない項目はどれなのか)は未知であることが多い。このような場合にDIF分析を行った結果として、50%以上の項目でDIFが生じている場合には、注意が必要である。

今後の課題

本研究では、パラメトリックな方法としてBILOG-MGによるDIFコマンド、ノンパラメトリックな方法としてSIBTESTを取り上げた。最初に述べたように、DIFを検出する方法には、ほかの方法もあり、今後はLordの χ 自乗法、尤度比による方法、Mantel-

Haenszel統計量の方法などを含めて、総合的な比較検討を行う必要がある。

引用文献

- 熊谷龍一 2003 語彙理解尺度のDIF分析の試みーBILOG-MGおよびSIBTESTを用いたDIF項目の発見ー日本心理学会第67回大会発表論文集, 365.
- Lord, F.M. 1980 *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Muraki, E. 2000 *RESGEN4* [Computer program].
- Roussos, L.A., & Stout, W.F. 1996 Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement*, 33, 215-230.
- Shealy, R., & Stout, W. 1993 A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, 58, 159-194.
- Stout, 1999 *DIFPACK: Dimensionality-based DIF/DBF package: SIBTEST, Poly-SIBTEST, Crossing SIBTEST, DIFSIM, DIFCOMP* [Computer program]. IL: Assessment System Corporation.
- 芝 祐順 1978 語彙理解尺度作成の試み 東京大学教育学部紀要, 17, 47-58.
- 芝 祐順・野口裕之 1982 語彙理解力尺度の研究 I 東京大学教育学部紀要, 22, 31-42.
- Thissen, D., Steinberg, L., & Wainer, H. 1993 "Detection of differential item functioning using the parameters of item response models." in P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum.
- 渡辺直登・野口裕之 1999 組織心理測定論 項目反応理論のフロンティア, 白桃書房
- Wilson, D. T., Wood, R., & Gibbons, R. 1985 *TESTFACT*. IL: Scientific Software International.
- Zimowski, M.F., Muraki, E., Mislevy, R. J., & Bock, R. D. 1996 *BILOG-MG* [Computer program]. IL: Scientific Software International.

謝 辞

本研究を実施するに当たって、語彙理解力尺度の利用を許可してくださいました東京大学名誉教授 芝 祐順教授に心から感謝いたします。

(2003年9月30日 受稿)

ABSTRACT

Comparing the Methods of Detecting DIF Using BILOG-MG and SIBTEST

Ryuichi KUMAGAI, Takafumi WAKITA

The purpose of this study was to compare the methods of detecting differential item functioning (DIF). We compared the parametric method which is executed by DIF command in BILOG-MG, and nonparametric method which is executed by SIBTEST. In the first study, we researched by simulation data. When the ratio with DIF item to all the items exceeds 50%, it was difficult to detect DIF correctly in both procedures. In second study, the real data, which is used scale for acquisition of Japanese word meanings, was researched. In both BILOG-MG and SIBTEST, the results of detection of DIF were approximately similar. However, estimated location parameters in BILOG-MG were very extreme, because latent distributional differences were very large between reference group and focal group.

Key words: differential item functioning, BILOG-MG, SIBTEST