項目パラメタドリフトの検出と原因の検討・共通尺度への影響

—そのドリフトはどこから来てどこへ行くのか—

寺尾尚大
大学入試センター 研究開発部

項目反応理論のモデルを適用し，潜在特性尺度 θ 上で表示される受検者個人の能力推定値 $\hat{\theta}_i$ を得たいとき，あらかじめ推定しておいた項目パラメタの値を既知として議論を進める。このとき，項目パラメタの推定時点と当該のテスト実施時点の間で値に大きな変化がないという前提に立っている。この前提は，項目パラメタの不変性 (item parameter invariance) と呼ばれる。

特定のテスト回に閉じた複数フォームの等化・対応づけなど，項目パラメタの推定時点と能力値の推定時点が同一であれば，項目パラメタの時間的不変性は特に問題にならない。一方，予備調査や過去の実施回における項目パラメタの推定結果を，時間をおいて本番のテスト回や他の実施回で能力値推定に利用する場合には，項目パラメタの時間的不変性が満たされていることを確認する必要がある。

具体的には，テスト項目をしばらくの間繰り返し出題するうちに，テスト項目の内容の変化やカリキュラムの改訂，項目漏洩や露出回数の増加などが起こると，推定時点と比べてテスト項目の困難度や識別力が変化する可能性がある。時点間で項目の困難度や識別力の値に標本変動の大きさを超える差異がみられる場合，項目パラメタドリフト (item parameter drift, IPD) が生じている。能力推定の前提である項目パラメタの不変性が阻害されている状況では，項目パラメタの値を固定して得られた能力推定値も妥当でないものになりうる。

本稿では，項目パラメタドリフトに関する理論的背景と分析事例を紹介し，項目パラメタドリフトの影響範囲や影響の大きさについて概観しながら，シンポジウムのテーマのひとつであるテスト項目の公開・非公開，およびその法的問題についての多面的な考察に寄与したい。

項目パラメタドリフトの概要と検出方法

まず，項目パラメタドリフトの現象について簡潔に説明する。例えば，2 パラメタロジスティックモデル (2PLM) の項目特性関数は，

$$P(\theta_i|a_j, b_j) = \frac{1}{1+\exp(-Da_j(\theta_i - b_j))}$$

のように表現され，項目特性に関するパラメタとして困難度 b_j，識別力 a_j を含む。運用上，Rasch モデルや 2PLM がよく適用されることも相まって，多くの先行研究では主として困難度 b_j・識別力 a_j のドリフトが取り上げられ，これに対して 3PLM の当て推量パラメタ c_j もあわせて検討対象としたものはごくわずかである。本稿でも，項目困難度および識別力パラメタのドリフトを主として取り上げることとする。なお，受検者の能力分布の変化 (construct shift) については，ここでは取り上げない。

項目パラメタドリフトは，教育測定学において長らく重要な研究テーマと認識されており，研究知見もすでに膨大に蓄積されている。項目パラメタドリフトの種類に初めて言及したであろう Mislevy (1982)，項目パラメタの時間的不変性を踏まえて能力に関する推論の必要性を提唱した Goldstein (1983) を皮切りにして，実際のテストデータに基づく分析，ドリフト検出手法の開発と評価，ドリフトの方向性や等化方法との関連の検討など，さまざまな観点から研究知見が提供されてきた。

実のところ，項目パラメタドリフトは，特異項目機能 (Differential Item Functioning, DIF) の特殊ケースとして理解されている。特異項目機能の分析では，受検者の所属集団の違いが，項目への応答や項目パラメタ，項目特性曲線に及ぼす影響の大きさを検討する。項目パラメタドリフトの分析では，所属集団の違いを取り上げる代わり

に，テストの実施回の違いを取り上げることになる。異なる実施回でテスト項目を繰り返し使用した場合，受検者が同じ項目に二度以上解答することは想定されていないため，受検者集団は互いに独立であると見なして分析を進めることとなる。

このことを利用して，項目パラメタドリフトの検出手法として，特異項目機能の分析手法がそのまま応用できるケースも多々ある。実際，古典的な分析としてよく用いられる Mantel-Haentzel 法やロジスティック回帰分析は，項目パラメタドリフトの検出の初手としても十分機能する。また，項目反応モデルを適用した分析としてよく用いられる尤度比検定や Lord の χ^2 統計量，項目特性曲線間の面積を求める Raju の方法なども有効である。

一方で，国内では項目反応理論に基づく尺度化事例が海外と比べて少ないこともあり，項目パラメタドリフトを表題に掲げて行われた研究は，並木・川端 (2019) に限定される。特異項目機能の分析に関する国内の精力的な研究動向 (e.g., 熊谷, 2012 ; 野口・熊谷・脇田・和田, 2007) と比べると，方法・実践の両面で項目パラメタの時間的不変性を精緻に捉える需要が国内になかったことも窺い知れる。

ドリフトの原因と方向性

項目パラメタドリフトが生じる主な原因には，主に二つのものが挙げられる。一つは，カリキュラムの改訂や項目内容の時間的変化など，測定内容や項目に関わるドリフトであり，もう一つは，項目の漏洩・露出回数の増加，不正行為など，テストセキュリティに深く関わるドリフトである (Haladyna & Rodriguez, 2013)。これとあわせて，特に困難度パラメタのドリフトの仕方が一方向 (unidirectional) であるか双方向 (bi-directional) であるかについても，注意深く検討を要する (Han, Wells, & Sireci, 2012)。

カリキュラムの改訂や項目内容の時間的な変化が原因となるドリフトは，基本的に項目の難易度を高める方向・低める方向の両方が考えられる。例えば，漢字テストにおいて「忖度」の読みを解答させる問題が出題された場合，漢字能力の分布には変化がない下では，メディア等で頻繁に取り上げられる前の困難度は高く，取り上げられた後の困難度は低くなっていることが想像される。また，項目識別力についても両方向のドリフトが考えられ，ホットトピックを含む項目では測定したい能力が高くなくても正答できる等の理由から識別力が低下することもあるだろうし，カリキュラムの改訂により一層尺度への関連が高まって識別力が向上することもあると思われる。

他方，本セッションのテーマに関連する項目漏洩・露出によるドリフトの場合，受検者同士でテスト項目の内容が共有され，項目困難度が低くなることが考えられる。また，能力水準の低い受検者が不適切に正答することとなり，項目識別力も低下する方向となる。項目パラメタのドリフト検出においては，突然易しくなったり，識別力が下がったりしている項目がないか，常に目を光らせておく必要がある。

ドリフトの影響範囲

項目パラメタに大きなドリフトが見られた場合，推定への影響が強く懸念される。ここでは，能力値と等化係数の二側面から，推定への影響を概観する。

第一に，項目パラメタドリフトの影響で，能力値の推定が歪められてしまうことが考えられる。意外なことに，ドリフトが能力推定値に及ぼす影響を検討した初期の研究では，能力値への影響が限定的であるという知見が説得的であった。Wells, Subkoviak & Serlin (2002) は，項目数・サンプルサイズ・ドリフトの種類などを操作し，シミュレーション実験により能力推定値への影響を検討したところ，その影響はごくわずかであったことを報告している。この研究では 2PLM を適用しているが，実際の運用において Rasch モデルや 2PLM などの倹約的なモデルを適用しているからこそ，多少項目パラメタの推定値がドリフトしてもその影響が吸収されるのかもしれない。一方，近年の研究では，Rasch モデルを用いた小サンプルサイズでの尺度化で，基準とするテスト回よりも項目困難度が高くなった場合に，能力値に基づく合否判定 (pass-fail decisions) への影響も指摘されている (Kopp & Jones, 2020)。

第二に，等化係数への影響が考えられる。もし，

異なるテスト回のチェイン等化 (chain equating) を実施する場合，項目パラメタドリフトが生じていると，等化係数の推定値にも影響を及ぼしうる。Han et al. (2012) は，困難度パラメタのドリフトの方向性と等化方法 (Mean-Mean 法，Mean-Sigma 法，Stocking-Lord 法) に着目して等化係数への影響を確認している。等化係数のうち傾き K (原著では A と表記) について興味深い結果が報告されており，Mean-Mean 法では中程度のドリフトが含まれていてもある程度頑健に推定できていたのに対し，Mean-Sigma 法や Stocking-Lord 法では項目パラメタドリフトの影響を少なからず受けていたことが示されている。チェイン等化を繰り返した場合には，ドリフトの影響から来る等化係数のズレも累積していく危険性がある点に留意が必要であると言える。項目パラメタドリフトの観点も考慮した等化方法の選択が，安定的な尺度構成の命運を握るだろう。

なお，項目パラメタドリフトは，得点の解釈基準にも大きな影響を及ぼしうる。特に，困難度パラメタの高低を解釈基準に反映させた standard setting を実施している場合には，ドリフトの影響を念入りに確認する必要がある。

分析事例

項目パラメタドリフトの分析例として，ここでは代表的な 3 つの事例を取り上げる。

Bock, Muraki, Pfeiffenberger (1988) は，実データを用いて項目パラメタドリフトの検討を行った先駆的な事例を提示している。College Board の物理テスト (the College Board Physics Achievement Test) を題材とし，いくつかの項目において困難度がドリフトしていたこと，識別力のドリフトが見られなかったことを報告している。さらに，項目パラメタドリフトの様相とカリキュラムの特徴を関連づけた分析も先進的である。

Wu, Li, Ng, & Zumbo (2006) は，1995 年・1999 年・2003 年に実施された TIMSS (Trends in International Mathematics and Science Study) のデータを用い，シンガポールと米国の中学 2 年生の数学を題材に，項目パラメタドリフトに関する分析を実施した。ロジスティック回帰分析を行い，合計得点を統制した上で，調査実施時期の主効果，合計得点と調査実施時期の交互作用の検討を行っている。結果として，どの項目においても，項目パラメタドリフトは確認されなかった。大規模な学力調査に基づく貴重な分析結果を提供している一方で，特にドリフトの原因については考察されていない。

Park, Lee, Xing (2016) は，TIMSS の中学 2 年生・数学のデータを分析対象としているが，1999 年・2003 年・2007 年の米国のデータのみを使用し，潜在クラスの発想を導入している点において，Wu et al. (2006) とは異なる視座からの分析結果を提供している。実際，特定の潜在クラスにおいて，3 回分の項目識別力パラメタが乱高下していた。合計得点ベースの分析だけでは見えてこないドリフトの検出に成功していると言える。こちらも，ドリフトの原因については考察されていない。

ドリフト検出時の対応

無視できない大きさの項目パラメタドリフトが生じた場合には，何らかの対応が求められる。ドリフト検出時の理想的な対応は，その項目を使用しないことである。項目パラメタを再推定するにせよ，一時的にその項目は出題の第一線から退くこととなる。項目パラメタドリフトの原因が項目漏洩であると推測されるときには，項目パラメタを再推定して利用することが難しく，実質的には使用中止の一択となる。

こうして見ると，項目パラメタドリフトを多様な方法で丁寧に検出しようとする割に，最終的な判断はとてもシンプルであるとも言える。Han et al. (2012) によれば，実際のところ事態はそれほど単純ではないようである。どの程度の大きさのドリフトが検出されれば出題を見合わせるのか，パラメタのドリフトが共通項目に生じていた場合の項目数との兼ね合い，ドリフトの原因や方向性に応じてどういった対応策を取るのか等，実務上悩みの種となりそうな事項が多くあるように思われる。他の資料も突き合わせてドリフトの原因を探ることはもちろん，そのテストで選抜や合否のカットポイントがある場合には，その付近でのドリフトのみに焦点化して検討するのも一案だ

と思われる。

最近登場した発想ではないが，online calibration の考え方に立つ研究者もいる。実際，Makransky & Glas (2010) では，過去の全テスト回のデータを累積的に用いて項目パラメタを毎回推定する方法が提案されている。テストの目的にもよるが，Fink, Born, Spoden, & Frey (2018) は，こうした大胆な方法を取った場合，同一の項目に対して毎回異なる推定値を用いることとなり，同一尺度上での能力値の比較を困難にするばかりか，ハイステークスなテストでは法的な問題に発展する危険性も指摘している。項目パラメタドリフトに敏感になりすぎて，共通尺度を常に浮動させることも適切ではなく，固定的な同一尺度上で能力について議論することの利点を失わない，バランスのよい判断が求められる。

おわりに

項目の非公開が比較的認められやすいテストであっても，項目パラメタの時間的不変性の仮定を満たせないケースがあったことは，極めて重要な知見である。項目を非公開にして統計的な管理が行える状況であっても安心はできず，不断のパラメタ管理が要求されることも，テスト関係者で共有すべき事項である。また，時間的変化も含めたさまざまな局外要因を取り除き，受検者の能力のみについて推論することの難しさを改めて痛感させられる。

最後に，本稿は，すでに項目反応理論を適用して安定的・継続的に運用されている試験・テストへのアンチテーゼではないことを付言しておきたい。国内において，項目反応理論を用いてパラメタの管理を実施されている方々の並々ならぬご尽力には，最大の敬意を表したい。また，項目パラメタドリフトに関する本稿の整理には，テスト項目の初出主義を促進する意図はまったくない。項目パラメタの時間的不変性について思いを致すことの重要性を話題提供するものである。

参考文献

Bock, R. D., Murakl, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, **25**(4), 275-285.

Fink, A., Born, S., Spoden, C., & Frey, A. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychological Test and Assessment Modeling*, **60**(3), 327-346.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement*, **20**(4), 369-377.

Han, K. T., Wells, C. S., & Sireci, S. G. (2012). The impact of multidirectional item parameter drift on IRT scaling coefficients and proficiency estimates. *Applied Measurement in Education*, **25**(2), 97-117.

Halaydyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

熊谷龍一 (2012). 統合的 DIF 検出方法の提案—“EasyDIF” の開発—. 心理学研究, **83**(1), 35-43.

Makransky, G., & Glas, C. A. W. (2014). An Automatic Online Calibration Design in Adaptive Testing. *Journal of Applied Testing Technology*, **11**(1), 1-20.

並木雄大・川端一光 (2019). 項目母数 Drift の存在が能力母数の推定に与える実質的影響. 日本テスト学会第 17 回大会発表論文抄録集, 138-141.

野口裕之・熊谷龍一・脇田貴文・和田晃子 (2007). 日本語 Can-do-statements における DIF 項目の検出. 日本言語テスト学会研究紀要, **10**, 106-118.
(Noguchi, H., Kumagai, R., Wakita, T., & Wada, A. (2007). Detection of DIF items in Japanese Can-do-statements. *JLTA Journal*, **10**, 106-118.)

Park, Y. S., Lee, Y.-S., Xing, K. (2016). Investigating the impact of item parameter drift for item response theory models with mixture distributions. *Frontiers in Psychology*, **7:255**, 1-17.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, **26**(1), 77-87.

Wu, A. D., Li, Z., Ng, S. L., & Zumbo, B. D. (2006). Investigating and comparing the item parameter drift in the mathematics anchor/trend items in TIMSS between Singapore and the United States. *Paper presented at the International Association for Educational Assessments*.

(terao@rd.dnc.ac.jp)