A Tutorial on Using Generative Models to Advance Psychological Science: Lessons From the Reliability Paradox Haines, N., Kvam, P. D., Irving, L., Smith, C. T., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., & Turner, B. M. (2025). A tutorial on using generative models to advance psychological science: Lessons from the reliability paradox. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000674

報告者：塚村 祐希

この論文は，ストループ課題や IAT のような代表的な行動課題において，集団平均の効果は頑健なのに，個人差指標の再検査信頼性が低いという「信頼性パラドックス」に対し，課題そのものの欠陥ではなく，データ分析の問題として捉え直し，生成モデルと階層ベイズモデルを用いることでこの（見かけ上の）パラドックスを大きく緩和できることを示したチュートリアルである。従来，ストループ効果などを「条件間の平均反応時間差」という単一の要約統計量で表し，その個人差と他の変数の相関を見る二段階アプローチが主流だった。しかし，平均差は反応時間分布のごく一部しか利用しておらず，個人ごとの推定値に測定誤差があることを事実上無視している。

そこで著者らは，理論に沿った生成モデルの構築を提案している。具体的には，反応時間データについて，正規分布モデル，対数正規分布モデル，シフト付き対数正規モデルを段階的に導入した。これを用いて再検査信頼性を推定する階層ベイズモデルを構築したところ，従来の二段階アプローチでは低い値となっていた　信頼性が，大幅に高い値に改善された。これは，従来の低い信頼性という見方が過度に悲観的であることを意味するとともに，生成モデルを用いることで，個人差の測定がより正確になることを示している。

Liu, Y., & Pek, J. (2024, February 8). Summed Versus Estimated Factor Scores: Considering Uncertainties When Using Observed Scores. *Psychological Methods*. Advance online publication. https://dx.doi.org/10.1037/met0000644

報告者：佐々木一洋

観測得点（例：総和得点や推定因子得点）はそのもととなる構成概念を反映すると想定されており，心理学研究で多くの用途がある。構成概念はしばしば潜在変数（LV）として操作化され，潜在変数測定モデルにおける観測変数との関係によって数学的に定義される。

この潜在変数に関して，近年では和得点（summed scores）と推定因子得点（estimated factor scores）の性質について議論が行われている（McNeish, 2022; McNeish & Wolf, 2020; Widaman & Revelle, 2022, 2023）。和得点は測定モデルにおける因子負荷の情報をほとんど利用しない点で「近似的（あるいは“粗い”や“厳密でない”）」手続きと呼ばれることがある（Gorsuch, 1983, p. 266; Horn, 1965）。しかし，因子負荷の情報を用いた推定因子得点の方が優れているかというと，どのような測定モデルであってもモデル誤差や標本誤差があるため，その影響の受けやすさも考慮する必要がある。本研究では，(a) 潜在得点の推定と個人の分類，および（b）潜在変数間の構造的関係の推定，という目的で和得点と推定因子得点（Bartlett 得点・回帰因子得点）の性能を比較検討する。特に，実用上の課題をよりよく反映するために，測定誤差だけでなく，モデル誤差と標本誤差を考慮に入れて評価する。

本稿では，まず和得点と推定因子得点，および共通因子モデルについて概説し，古典的テスト理論（CTT）に基づいて両者の信頼性を比較できることについて述べた。そのうえで，モンテカルロ・シミュレーションによって，信頼性・標本サイズ・モデル誤差の異なる条件下で各得点がどのように振る舞うかを検討した。さらに，(a)個人の分類，(b)潜在変数間の構造的関係の推定のそれぞれについて，実データの例を示した。

シミュレーションの結果，推定因子得点は，測定モデルがよく適合し標本が十分大きい条件では信頼性・精度で有利であり，特に回帰因子得点は理論上最も高い信頼性を示した。一方でモデル誤差や標本サイズの小ささは推定因子得点の性能を著しく低下させ，単純な和得点が相対的に頑健であった。また，潜在変数間の構造的関係の推定については，SEM は大標本かつ妥当なモデルのとき最も優れていたが，小標本やモデル誤差が大きいと不安定になり，和得点を用いる回帰モデルがより頑健であった。

このように，測定モデルがよく適合し標本が十分大きければ推定因子得点が望ましいものの，モデル適合に疑義がある場合や，サンプル数が少なく，特に収束問題が生じる恐れがある場合は，和得点が現実的であるといえる。どちらを選択するかを考えるうえでは，信頼性（測定誤差）だけでなく標本誤差とモデル誤差も合わせて評価することが望ましい。

Player, L., Hughes, R., Mitev, K., Whitmarsh, L., Demski, C., Nash, N., Papakonstantinou, T., & Wilson, M. (2025). The use of large language models for qualitative research: The Deep Computational Text Analyser (DECOTA).*Psychological Methods.* Advance online publication. https://doi.org/10.1037/met0000753

報告者：伊藤　黎

本論文で提案された手法 DECOTA では、テキストデータを複数テーマに分類することが目的とされている。まず、各文を一つの文書として見立てた上で構造化トピックモデリング（STM）を用いて、トピック分類がなされた。この各トピックが各々コードとして見立てられ、トピック内の特徴語と文サンプルをもとに、そのコード名をファインチューニングされた LLM が命名する。その次にコード名を BERT により分散表現に変換し、コサイン類似度をもとにクラスタリングが行われた。この工程で形成されたクラスタをテーマとして見立て、ファインチューニングされた二つ目の LLM がテーマ名を命名する。この命名されたテーマ群が最終的なアウトプットとして、テキストデータのテーマ分類結果となる。

本手法の検証として、四つのテキストデータセットに対し、それぞれの元論文で帰納的テーマ分析がなされた結果、本論文の著者が帰納的テーマ分析を行った結果、そして DECOTA によって得られたテーマ分類結果を比較する三角測定が行われた。人間と DECOTA のコーディングやテーマ分類には高いレベルの一致率が見られ、テキストデータを質的に分析する高速かつ信頼性の高い手法として、DECOTA の有用性が示された。

ただし、本手法は同一の質問に対する回答文といった構造のテキストデータのみに対応している点、コーディングが古典的自然言語手法である STM に依存しているため意味的・感情的解釈を必要とするコーディングには不向きである点、全体の分類を示すだけで少数の意見を抽出できない点などの制約に注意する必要がある。

Siepe, B. S., Bartoš, F., Morris, T. P., Boulesteix, A.-L., Heck, D. W., & Pawel, S. (2024). Simulation studies for methodological research in psychology: A standardized template for planning, preregistration, and reporting. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000695

報告者：加藤剛

シミュレーション研究は，心理学における統計手法の性能評価に広く用いられている一方で，計画・実施・報告の質には大きなばらつきがある。また，設計上の自由度が高く，条件設定や評価指標の選択といった研究者の裁量が，結果の解釈や選択的報告に影響しうる。本論文では，2021～2022 年に Psychological Methods，Behavior Research Methods，Multivariate Behavioral Research に掲載された 321 本の論文をレビューし，そのうちシミュレーション研究を含む 100 本（31.2%）を対象に，報告実践の現状と問題点を整理した。

レビューの結果，報告すべき情報の欠如が複数の点で確認された。まず，シミュレーションの反復回数について，その根拠を明示する研究は少数（8.0%）にとどまり，性能指標に付随するモンテカルロ不確実性（e.g.，モンテカルロ標準誤差）を報告しない研究も多かった（77.0%）。さらに，コードが公開されていない研究が相当数あり（64.0%），計算環境について全く報告しない研究も少なくなかった（24.0%）。加えて，推定対象の定義が曖昧なまま比較が行われていたり，収束判定に関する報告が不十分だったりするなど，方法の妥当性や結果の解釈に関わる記述の不足も指摘されている。

これらの課題に対処するために，本論文では Morris et al.（2019）の ADEMP（Aims, Data-generating mechanism, Estimands and other targets, Methods, Performance measures）を心理学向けに整理し，計画・事前登録・報告を標準化するための ADEMP-PreReg テンプレートを提案している。このテンプレートは，目的の明確化，データ生成メカニズムの仕様化，推定対象（estimand）の定義，比較する方法の選択，性能指標の決定といった要素を順序立てて記述できるよう設計されており，代表的な性能指標の定義，モンテカルロ標準誤差，目標の精度を達成するための反復回数の計算法も含む。

最後に，設計・分析上の選択に根拠を与えること，ADEMP のような標準化された枠組みを用いること，モンテカルロ不確実性を明示すること，目標の精度に基づいて反復回数を決めること，可能なら事前登録を含む研究プロトコルを整備すること，FAIR 原則に沿った形でコードを公開すること，ソフトウェアや計算環境を詳細に報告することを推奨している。また，単一のシミュレーション研究の一般化可能性には限界がある点を明確にし，限られた計算資源の下で，反復回数（精度）と条件設定の網羅性（一般化可能性）のバランスを意識して設計する必要があると指摘する。ジャーナルや査読者も，コードの共有の要求や再現性のチェック等を通じて，報告実践の改善を促進しうると結論づけている。

Orsoni, M., Benassi, M., & Scutari, M. (2025). Information theory, machine learning, and Bayesian networks in the analysis of dichotomous and Likert responses for questionnaire psychometric validation. *Psychological methods*.

報告者：今泉智子

心理学では，理論と経験的データが十分に結びついていないことが再現性の危機の一因となっている（Oberauer & Lewandowsky, 2019）．特に構成概念の因果構造をどう扱うかが課題となっている．質問紙は構成概念に対応する症状・行動を測定するための中心的ツールであり，質問紙の項目は構成概念と直接関係し，その内容領域を代表している必要がある．従来の潜在因子モデルとしてEFA，CFA，SEMが挙げられるが，これらは正規性を前提としていること，局所独立の仮定が強いこと，主観的判断（因子数，回転など）が混入することなど限界があり，構成概念の因果的理解を妨げる可能性があった．近年，心理学的構成概念を「症状や行動の因果ネットワーク」として捉えるアプローチが用いられつつある．これは例えば，うつ病は潜在因子が症状を生むのではなく，症状同士の相互作用からうつ病という状態が形成されるという考え方である．このアプローチはネットワークのパターンからグループを判別することを可能にする．

そこで本研究では，(1) 確率分布の類似度を測るJSダイバージェンスを用いて2つのグループの回答がどれだけ異なるかを定量化し，グループの識別力の高い項目を選抜，(2) 様々なMLアルゴリズム（教師あり学習）を適用，比較し最も正確に被験者を分類できるモデルを選択，(3) ベイジアンネットワークを用いて，変数間の依存関係を解明という3つのプロセスで質問紙の妥当性を検証する枠組みを提案した．提案手法の検証として回答の傾向が異なる2つのグループ間において2値データと順序データの場合でのシミュレーションとTMAS（顕在的不安尺度，2値・50項目）データを用いた実データ分析を行った．その結果，質問紙項目をJSダイバージェンスの大きさをもとに大幅に削減してもMLモデルの分類精度は高いパフォーマンスを維持すること，そして，学習されたネットワーク（logBFでの構造評価ではTabusearchでのDAG作成がPCStableと比較してよりよい構造をつくるこがわかった）を用いて項目間の関係を確率的に分析できることを確かめられた．

JSダイバージェンスによる項目選択とMLによる分類精度の検証は，質問紙の基準関連妥当性と内容的妥当性を直接的に評価するものである．そして，回答者の負担軽減とデータの仮定の回避，解釈における主観性の排除を達成する．さらにBNを用いることで症状（項目）間の確率的な相互作用を詳細に理解でき，既存の理論の確認や新しい仮説の生成にもつながる．

今後は，意図しない構成概念との相関を避けるための指標をDifferential Reliability Index；DRIの概念を拡張し開発することで弁別的妥当性の評価を組み込むことが展望として挙げられる．※本研究では，心理学的構成概念は相互に因果的に影響し合うシステムであること，グループにおける差（集団差）は既知であること，質問紙の項目削減が目標であることを前提としている．

A Family of Sequential Item Response Models for Multiple-Choice, Multiple-Attempt Test Items. *Psychometrika*, 90 (1), 331-357. https://doi.org/10.1017/psy.2024.18

報告者：南元　篤史

本研究では，1 つの項目に対し複数回に渡って解答することが可能な，多試行形式と呼ばれる試験に焦点を当てる。試行回数は上限を定めることもできるが，究極の場合は正解するまで解答する（answer-until-correct; AUC）形式となる。このような解答形式について先行研究では，学習効果や信頼性の向上をもたらす可能性が示されている。多試行形式に対応する項目反応理論（item response theory; IRT）モデルとして逐次項目反応理論（sequential IRT; SIRT）モデルが存在するが，既存の SIRT モデルは当て推量による正答の可能性を考慮していない。多肢選択式項目を用いた多試行形式では，誤答の度に可能な選択肢が減っていき，特に後の試行では当て推量の可能性が無視できないものとなる。そこで本研究では，当て推量の可能性を考慮した，多肢選択式多試行形式項目のための SIRT（SIRT-multiple-choice multiple-attempt; SIRT-MM）モデルを提案した。離散選択理論を援用して SIRT-MM の一般的な枠組みを定義した後，最も簡単な場合として，すべての不正解選択肢が均質な場合のモデル

$$P(X=u|\theta)=\frac{(K-1)!\left(1+K exp\left(a(\theta-b)\right)\right)}{(K-u)!\prod_{k=1}^{u}\left(K-k+1+K\exp\left(a(\theta-b)\right)\right)}$$

を導出した。ここでXは正解までに要した試行回数，Kは選択肢数，a, bは項目パラメタ，θは個人パラメタ（能力）である。さらに多様な項目を表現するために，不正解選択肢の不均一性を表す困難度シフトパラメタγ_uを導入した一般のモデル

$$P(X=u|\theta)=\frac{(K-1)!\left(1+K exp\left(a(\theta-b+\gamma_u)\right)\right)}{(K-u)!\prod_{k=1}^{u}\left(K-k+1+K\exp\left(a(\theta-b+\gamma_u)\right)\right)}$$

を導出した。これらのモデルから生成したシミュレーションデータに対し，一般的な多値型 IRT モデルである段階反応モデルや名義反応モデルを含む複数のモデルを AIC や BIC で比較したところ，多くの場合に生成モデルである SIRT-MM モデルが正しく支持された。また，項目パラメタ復元のシミュレーションでは，γパラメタを含まないモデルでは N=500 程度，含むモデルでは N=1000～2000 程度以上のサンプルサイズが望ましいことが分かった。個人パラメタ推定値は，多試行形式と SIRT-MM モデルを採用することで単試行形式よりも高精度となることがシミュレーションと実データ分析を通じて示された。

Yaremych, H. E., Preacher, K. J., & Hedeker, D. (2023). Centering categorical predictors in multilevel models: Best practices and interpretation. *Psychological Methods*, *28*(3), 613–630. https://doi.org/10.1037/met0000434

報告者：加藤剛

マルチレベルモデル（MLM）におけるセンタリング方法の選択は，パラメータの推定と解釈にとって重要である。しかし，センタリングに関する方法論的議論は連続変数に集中しており，カテゴリカル変数をセンタリングすべきかどうか，どのように解釈すべきかについては十分に検討されてこなかった。応用研究のレビューでは，カテゴリカル変数のセンタリングを適切に行う研究は少なく，センタリング方法の選択根拠や結果の解釈の記述が不十分であることが示された。統計的には連続変数とカテゴリカル変数は同様に扱えるが，カテゴリカル変数の回帰係数の解釈は直感的ではない。本論文では，特にランダム切片固定傾きモデルの下でのカテゴリカル変数のセンタリング方法と回帰係数の解釈について論じる。

カテゴリカル変数の効果を正しく推定・解釈するには，コーディング方法とセンタリング方法の適切な選択が必要である。コーディング方法としてはダミーコード，対比コード，効果コードが取り上げられ，期待値の導出により，各コーディングの下での傾きとグループ平均差の対応関係が明らかにされた。センタリング方法としては，センタリングなし（UN），全体平均中心化（CGM），クラスタ平均中心化（CWC）が検討された。UN モデルと CGM モデルは等価であり切片の解釈のみが異なるが，いずれも傾きの推定値はクラスタ内効果とクラスタ間効果の解釈不能な混合となる。一方，CWC 予測変数とクラスタ平均を同時に含む CWC(M)モデルでは，両者が無相関となるため，クラスタ内効果とクラスタ間効果を独立に解釈できる。また，UN 予測変数とクラスタ平均を含む UN(M)モデルは CWC(M)モデルと等価だが，レベル 2 の傾きはクラスタ間効果ではなく文脈効果として解釈される。

実証例として，スコットランドの学業成績データを用いて，親の教育状況（4 カテゴリ）が子どもの成績に与える影響が分析された。UN モデルの傾きの推定値はクラスタ内効果とクラスタ間効果のいずれとも異なり，この例ではクラスタ内効果を過大推定していた。CWC(M)モデルによる分析では，同じ学校内での親の教育の効果と，学校間での親の教育状況の構成の効果が区別され，後者の方が大きいことが示された.

実践上の考慮事項として，共変量のセンタリング，交互作用項の推定，有意性検定，潜在変数を用いたマルチレベル SEM との関係についても議論された。特に，レベル 1 のカテゴリカル共変量についても，関心のある予測変数と同様にセンタリングすべきであり，交互作用項を含める場合は CWC によりクラスタ内効果を分離してから推定する必要がある。今後の研究課題として，ランダム傾きを含むモデルや，従属変数が連続変数以外である場合（e.g., 二値データ，カウントデータ）への拡張が挙げられる。