大規模言語モデルの翻訳を評価する大規模共同研究：心理尺度の人手-機械翻訳間の24比較

A big team study evaluating translations with large language models: Comparisons between human and machine translations across 24 psychological measures

山田祐樹[1]*・小杉考司[2]・国里愛彦[2]・分寺杏介[3]・後藤崇志[4]・橋本泰央[5]・工藤大介[6]・李禕飛[7,8]・眞嶋良全[9]・向井智哉[10]・野村竜也[11]・小口真奈[12]・七條花恋[1]・下司忠大[13]・高松礼奈[14]・竹橋洋毅[15]・竹下昌志[16]・浅野良輔[17]・福田実奈[18]・古谷嘉一郎[19]・日道俊之[20]・平野寛樹[20]・五十嵐祐[16]・伊藤雅隆[21]・香川璃奈[22]・神野雄[23]・加藤弘通[24]・古村健太郎[25]・宮川裕基[26]・水野君平[27]・村浦新之助[28]・新谷優[29]・西村多久磨[30]・尾崎由佳[31]・佐藤秀樹[32]・佐藤奈月[24]・嶋大樹[26]・瀧川諒子[33]・田中勝則[34]・塚本早織[14]・山崎茜[35]・楊帆[36]・三浦麻子[4]*

[1]九州大学・[2]専修大学・[3]神戸大学・[4]大阪大学・[5]東北文教大学・[6]東北学院大学・[7]東京都立大学・[8]帝京大学・[9]北星学園大学・[10]福山大学・[11]龍谷大学・[12]沖縄科学技術大学院大学・[13]立正大学・[14]愛知学院大学・[15]奈良女子大学・[16]名古屋大学・[17]久留米大学・[18]京都外国語大学・[19]関西大学・[20]高知工科大学・[21]福島大学・[22]筑波大学・[23]西武文理大学・[24]北海道大学・[25]弘前大学・[26]追手門学院大学・[27]北海道教育大学旭川校・[28]上越教育大学・[29]法政大学・[30]東京理科大学・[31]東洋大学・[32]福島県立医科大学・[33]北陸先端科学技術大学院大学・[34]北海学園大学・[35]広島大学・[36]早稲田大学

*連絡著者：yamadayuk@gmail.com

欄概略題：大規模言語モデルによる尺度翻訳

利益相反：全ての著者について報告すべき利益相反はありません。

アブストラクト

大規模言語モデル (LLM) は心理学研究の多側面にて活用されているが，心理尺度の翻訳において人手翻訳と同等の実用性をもつかは未検証である。このManyScalesプロジェクトは，LLM翻訳の妥当性と実用性を多面的に評価し，人手翻訳との比較を行うことを目的とする。本研究は日本国内での大規模共同研究 (43名，36機関) により実施され，24種類の英語版心理尺度を対象とする。各尺度について，Rパッケージ*LLMTranslate*を用いたLLM翻訳版と，翻訳者による人手翻訳版を作成し，いずれも統一した手続きで逆翻訳を行う。両翻訳版は，(a) 専門家による意味的忠実性・自然さ・文化的妥当性の評価，(b)一般回答者による理解しやすさと自然さの評価，(c) 心理測定学的分析 (因子構造・因子得点・測定不変性・関連係数など) の観点から比較する。さらに探索的に，埋め込み表現に基づくコサイン類似度を算出し，原版項目との意味的距離を検討する。本研究により，LLM翻訳が心理尺度翻訳にどの程度活用可能であるか，また人手翻訳との差異がどの観点に表れるかを明らかにし，尺度翻訳プロセスの可視化・標準化への貢献を目指す。

Abstract

Large Language Models (LLMs) are widely used in many aspects of psychological research, but their applicability in translating psychological measures to match human translation remains unverified. This ManyScales project aims to evaluate the validity and applicability of LLM translations from multiple perspectives and compare them with human translations. This study is a large-scale collaborative endeavour within Japan (43 researchers, 36 institutions), focusing on 24 English-language psychological scales. For each scale, an LLM-translated version was developed using the R package *LLMTranslate*, alongside a human-translated version by professional researchers. Both versions are back-translated using a common procedure. Both translation versions will be compared from the viewpoint of (a) expert evaluation of semantic fidelity, naturalness, and cultural validity; (b) lay participants' assessment of understandability and naturalness; and (c) psychometric analysis (factor structure, factor scores, measurement invariance, some coefficients, etc.). Furthermore, we will explore cosine similarity based on embedded representations to examine semantic distance from the original items. This research aims to reveal the extent to which LLM translation can be used for psychometric scale translation and where differences from human translation manifest, thereby contributing to the visualization and standardization of the scale translation process.

キーワード

心理尺度，大規模言語モデル，機械翻訳，人手翻訳，ビッグチームサイエンス，翻訳精度

Keywords

psychological scale, Large Language Models, machine translation, human translatio
n, big team science, translation fidelity

イントロダクション

人工知能 (Artificial Intelligence: AI) や機械学習は，社会生活における非常に広範な領域での基盤技術として定着しつつある (Maslej et al., 2025)。我々の日常生活において，AI技術はもはや珍しいものではなく，多くの場合とくに意識されることもなく利用されている。例えば，NetflixやSpotifyのようなストリーミングサービスでは，AI (特に機械学習アルゴリズム) を用いて個人の視聴・聴取履歴を分析し，コンテンツを推薦する。また，AppleのSiriやGoogleアシスタントに代表されるスマートフォンの音声アシスタントでは，AIを用いてスケジュールの確認，天気予報の検索，簡単な質問への応答といった日常的なタスクを実行している。さらに，電子メールのスパムフィルタリングもAI (機械学習) の普及した応用例である。Gmailのような主要なプロバイダーは，機械学習アルゴリズムを用いて受信トレイに入るスパムメールを99.9%以上ブロックしていると発表されている (Google, 2025)。このように，AIはレコメンデーション，音声認識，気象予測，パターン分類といった形で，すでに社会的インフラストラクチャーの一部として機能している。

こうしたAI技術の普及はより日常的な研究活動においても急速に一般化している。特に大規模言語モデル (Large Language Model: LLM) は，研究の生産性を向上させるツールとして多用されている。GitHub Copilotのようなツールは，データ分析時のRやPythonのコード生成を支援する。これにより，従来は時間のかかっていたデータの前処理や可視化のスクリプト作成が効率化されている。同様に，研究者はElicit (https://elicit.com/)，Consensus (https://consensus.app/)，Scite.ai (https://scite.ai/) といったAIを活用した検索エンジンやレビュー支援ツールを用いて，膨大な文献レビューを効率化し，研究の着想を得ている。これらのツールは，単なるキーワード検索とは異なり，研究課題に基づいた論文の検索，要約の抽出，あるいは論文間の引用関係の文脈的な分析までもを可能にしている。さらに，社会科学の研究プロセスにおいては，LLMは質的データのコーディング (Friese, 2025) や，データアノテーション，心理学実験で使用するシナリオ，単語集，あるいは画像・動画刺激の生成といった，より専門的な用途にも用いられている。

心理学とその関連領域においては，LLMは研究支援ツールであるだけに留まらず，人間の認知や行動を模倣・代替する「対象」としても扱われ始めている (Binz, Alaniz et al., 2025)。特に，LLMを心理学実験の「被験者」として扱うアプローチと，対人支援を行う「臨床家」として扱うアプローチの二点においてこれが顕著である。

第一に，LLMを人間の代替被験者として利用する可能性が検討されている。Dillion et al. (2023) は，LLMが道徳的判断や経済的ゲームといった課題において人間と類似した反応を示すことから，予備的なデータ収集や仮説生成において人間の参加者を補完するための「シリコン・サンプリング (silicon sampling)」としての有用性を提唱した (e.g., Argyle et al., 2023)。Huang et al., (2024) はLLMで生成した各エージェントに異なるパーソナリティ特性を割り当ててパーソナリティ質問紙への回答生成を求め，エージェントが人間のパーソナ

リティ特性を効果的にシミュレートできることを示した。また，Binz & Schulz (2023) は，GPT-3に対して認知心理学の実験課題を実施し，意思決定における学習パターンや情報探索行動が人間と類似することを示した。ただし，因果推論においては人間と異なり困難であったことも報告している。さらに，Mei et al. (2024) は，最後通牒ゲームや公共財ゲームを含む一連の行動経済学ゲームとビッグファイブ尺度を用いた行動的チューリングテストを実施し，GPT-4の行動特性が，50カ国以上の人間からなる大規模サンプルのパフォーマンスと大きく異ならないことを示した。この知見は，LLMが認知課題や社会行動といった従来の心理実験における人間の反応パターンをほぼ忠実に再現できることを示唆している。

最近では，BinzとSchulzは，160の心理学実験から得られた1000万試行以上にも及ぶ人間の選択行動についての大規模データセットであるPsych-101 (https://huggingface.co/datasets/marcelbinz/Psych-101) に基づき，Llama 3.1ベースで大規模に学習させた基盤モデル「Centaur」を開発し，未知の実験パラダイムにおける人間の行動予測において既存の認知モデルを上回る精度を示したことを報告している (Binz, Akata et al., 2025)。これは，心理学において各個発展してきた様々な認知モデルが，データ駆動型の統一理論に置き換えられてしまう可能性を示し，衝撃を与えた。しかし，Centaurは単に行動予測が優れているだけで認知メカニズムの説明ができていない (Bowers et al., 2025) どころか認知のモデルですらない (Orr et al., 2025) といった批判や，実験課題の種類によっては (Namazova et al., 2025)，あるいは実験課題の設定を僅かに調整しただけでも (Schröder et al., 2025) 人間の反応パターンと大幅に乖離してしまうという指摘など，その性能や貢献を疑問視する意見が多い。たとえテスト時に教示や問題文といった選択肢の記号以外の全てを削除した場合であってもCentaurが高いパフォーマンスを示したことから，単なるパターンマッチングや過剰適合のようなことが起きているのではという指摘もある (Liu & Ding, 2025)。付け加えるなら，そもそもファインチューニング前のLLM学習時に既に各課題についてリーケージがあった可能性も考えられる。このように，LLMが真に人間の認知モデルとなり得るかについては今まさにホットな議論が行われている状況である。

第二に，LLMをメンタルヘルス領域における臨床家やカウンセラーとして応用する試みも行われている。Elyoseph et al. (2023) は，ChatGPTが感情への気づき段階尺度 (LEAS) において一般成人男性の平均を上回る得点を示し，LLMの感情的コンピテンスが高いことを示唆した。より実践的な文脈においては，Ayers et al. (2023) が，患者からの医療相談に対する回答の質と共感性を医師とチャットボット (ChatGPT) で比較検討している。その結果，チャットボットの回答は医師と比較して「質が高い」と評価される割合が圧倒的に高く (医師：22.1% vs. チャットボット：78.5%)，また「共感的である」と評価される割合についても同様であることが示された (医師：4.6% vs. チャットボット：45.1%)。これらの知見は，LLMが言語的なコミュニケーションにおいて，人間と同等あるいはそれ以上の対人支援能力を発揮しうることを示唆している。さらに最近のランダム化比較試験による臨床研究では，

生成AI (Falcon-7BおよびLLaMA-2) を用いたチャットボットによる対話が，未使用群よりもうつ病や不安症状を有意に大きく軽減させることが示された (Heinz et al., 2025)。

こうしたAIの対話能力は，治療者としてだけでなく，治療者を訓練するための「クライエント・サロゲート (模擬患者)」として活用できる可能性を示す。Teixeira da Silva & Yamada (2024) は，LLMを正式なカウンセリング実践に直接適用する際には多様な利点とリスクが並存することを指摘しつつ，教育的文脈において患者役として用いる可能性にも触れている。これと関連して，Cross et al. (2025) は，医学生がChatGPTを模擬患者として問診練習を行うロールプレイを実施し，LLMによる模擬患者の課題として，プロンプト設計の難しさや非言語的手がかりの欠如などがあることを報告している。また，心理学の学生を対象にLLM模擬患者との臨床面接シミュレーションを行った研究では，学生の不安を含むネガティブ感情が有意に減少し，知識や技能に関する自己評価が向上した (Sanz et al. 2025)。参加者はLLMとの対話が実践的な準備として有益であると報告していた。このように，LLMは安全かつ効果的な教育ツールとしても機能しうるとして，さらなる検討が進んでいる。

心理学研究の参加者，カウンセラー，患者を模倣するようなLLMの能力は，心理学研究で用いられる様々なマテリアルの作成においても発揮されており，それは心理概念の測定を行う質問紙の開発においても同様である。質問紙の開発・改訂・翻訳には時間と専門的判断が求められ，時にはインタビューや複数の専門家によるブレーンストーミングといった質的研究による手法も介在し，研究者の負担も大きい (e.g., Stefana et al., 2024)。この状況に対して，自然言語処理 (Natural Lanuage Processing: NLP) やLLMを活用し，心理概念の測定や尺度開発の自動化や効率化する試みが進んでいる。例えば，Kjel et al. (2019) では，自由記述テキストをNLPで意味分析した方法と，従来の心理尺度における自己報告 (回答者が項目を読み，項目内容に対する当てはまり度を報告する方法) を比較した。その結果，両者は同等，あるいは前者の方が高い信頼性と妥当性を示し，自然言語に基づく測定の有効性が示唆された。さらに，LLMの活用は尺度開発プロセスにおいても及んでいる。Symeonaki et al. (2024) は，特定の集団に対する態度を測定する尺度項目を用い，人手による項目分類とLLMによる意味関連性に基づく分類を比較した。具体的には，75名の参加者 (学部生および大学院生) による項目分類と，LLM による意味や関連性に基づく項目分類を並行して分析し，両者の一致度を評価した。その結果，人間と LLM の間で高い項目分類の一致度が示された。この結果から，尺度の項目評価にLLMを活用できる可能性が示されている。加えて，複数の研究において，ChatGPTを用いて尺度項目を自動生成し，質の高い項目の生成が可能であることが報告されている (e.g., Götz et al., 2024; Maertens et al., 2024; 佐々木・豊田, 2024)。これらの知見から，LLMは項目評価だけでなく，尺度項目の生成プロセスそのものを支援し，部分的に代替しうることが示唆される。

このように，2024年前後より，心理尺度開発においてLLMを活用した項目生成や項目評価の有効性が報告されるようになってきた。しかし，尺度開発プロセスが自動化・効率化されたとしても，その使用に際しては言語的・文化的制約が依然として存在する。多くの尺度

は特定の言語圏，とりわけ英語圏を中心に作成されており (e.g., Zhang et al., 2024)，他言語・多文化への適用には，翻訳と文化適応が不可欠である (Massé et al., 2025; Salama-Younes et al., 2009)。心理尺度の翻訳ガイドラインは多数提案されているものの，それらのガイドラインの中で共通して記載されるような統一的・絶対的な指針までは示されていない (Cruchinho et al., 2024; Epstein et al., 2015)。その中でも，尺度翻訳の標準的手続きとして，原版から該当言語への「順翻訳」，順翻訳を原版の言語に翻訳し直す「逆翻訳」，さらに逆翻訳と原版の意味的乖離を確認し，適宜翻訳を修正する「調和」などの手続きを含む逆翻訳法は広く推奨されてきた (e.g., Brislin, 1986; Harkness, 2003)。

しかし，実務上は逆翻訳法にも課題は多い。Granas et al. (2014) は，患者の薬に関する信念尺度 (Belief about Medicines Questionnaire) の翻訳版 (ノルウェー語・スウェーデン語・デンマーク語) を比較し，原版の表現が翻訳後には文化的背景や日常的用法の違いによって異なる解釈を生む場合があることを示した。つまり，翻訳において語彙レベルでの一致 (意味的等価性：semantic equivalence) が保たれていても，受け手の理解や解釈が原版と一致するとは限らない。このような課題に対し，Ozolins et al. (2020) は，逆翻訳が字義的一致の確認に偏る傾向を指摘し，翻訳者と研究者が測定概念や研究目的をふまえて訳語選択するプロセスの重要性をあげている。さらに，訳語選択の根拠が論文中で十分に説明されていない点も課題として指摘している。この傾向は，翻訳版も含む患者・医療従事者中心性尺度 (Patient-Practitioner Orientation Scale) のスコーピングレビューにおいても報告されている (Werner et al., 2025)。その結果として，翻訳プロセスの透明性と再現性が損なわれ，国際比較研究の妥当性に影響を及ぼす可能性がある。すなわち，尺度翻訳に求められるのは単なる語彙一致ではなく，原版と同様の理解や心理的反応が生じる語用論的等価性 (pragmatic equivalence) にある。しかし，このような語用論的等価性の判断は，翻訳者や研究者の言語感覚や理論的理解に依存し，客観的な検証・透明化には限界が生じる。

逆翻訳法を用いた心理尺度の翻訳は，日本語話者を対象に研究を実施することの多い日本の心理学研究において頻繁に行われている。日本国内の学会が発行する「心理学研究」「パーソナリティ研究」の2誌について，2019年度から2023年度に公刊された論文をレビューした[1]。具体的に対象としたのは「心理学研究」第90巻1号から第94巻6号および「パーソナリティ研究」第28巻1号から第32巻3号に掲載された論文であり，J-STAGEで公開されているpdfを参照した。本レビューにおける「心理尺度」は「尺度」あるいは英語で”scale”，“questionnaire”等と明記されているものに限定せず，単数・複数の項目に対して参加者・調査回答者自身が自己報告で回答する，あるいは保護者等が子ども等の他者の言動について回答するものであり，対象者の心理状態を数量的な指標により推定するために用いられているものを対象とした。実験刺激や場面想定法等で用いられるシナリオの日本語翻訳は対象としていない。また，明確に日本語訳・邦訳・和訳したことが記述されているものに加え，海外文

[1] このレビューにおける情報収集と分析は後藤崇志が主に行った。分析に使用したデータは補足情報としてOSFにて公開されている (https://osf.io/hn4fr/files/gc86k)。

献で使用されている心理尺度を引用して使用しており，当該文献で日本語話者が調査対象者となっておらず日本語翻訳版が存在しなかったであろうと推測されるものも，研究の中で日本語翻訳が行われたものと推測して含めている。他方で，海外文献や海外研究で作成された心理尺度を参考に作成したと記述されていたものについては，日本語翻訳以上の変更が加えられていると推測して含めていない。日本語以外の言語への心理尺度の翻訳を行っている研究も含めていない。また，本レビューでは逆翻訳の有無，等価性の確認の有無，認知デブリーフィングや表現の調整・修正の有無に関しては，明確に記述が見られたか否かという観点で判断した。したがって，実際には行われていたものの論文上には記述されていなかったものがある可能性もあることは留意する必要がある。

対象となった407報の論文のうち，71報 (17.4%) において心理尺度の日本語翻訳版が作成されていた (1報で複数の日本語翻訳版を作成していたものもあるため，作成されていた日本語翻訳版の心理尺度は80件であった)。国内誌のレビューで見られた80件の心理尺度の日本語翻訳版作成過程の記述を見る限り，55件 (68.8%) の心理尺度で逆翻訳法が用いられていた。46件 (57.5%) では，逆翻訳を行った後に等価性の確認を行われており，そのうちの39件は原版の心理尺度を作成した著者も等価性の確認に関与していたとする記述が見られた。しかしながら，逆翻訳法を行う過程においてどのような点に留意して翻訳や修正が行われたかは明確に記述されていないものが多い。たとえば，30件 (37.5%) の尺度では複数者が順翻訳を行っていたことが明記されていたが，複数の翻訳版をどのように統合したかはほとんど明記されていない。また，順翻訳の後に表現の調整や表現理解に関する予備調査を行ったものが16件 (20.0%)，翻訳を終えた後に認知デブリーフィングや予備調査を行い，表現の修正を行っているものが12件 (15.0%) 見られたが，いずれも語の認識の揺れを考慮した教示の追記を行ったことを明記していたものや、文化的に馴染みのない事物を日本で馴染みのあるものに置き換えていたものが数件見られたまでであり，ほとんどの論文では明記されていない。少なくとも心理尺度の日本語翻訳において，原版と翻訳版の等価性を担保し，妥当な心理尺度を作成しようとするために，各研究で具体的にどのような手続きが取られているのかについて，逆翻訳法が用いられているということ以上のことは明確には報告されていない現状にある。心理尺度の翻訳版を作成するプロセスの透明性が確保されていないことは，逆翻訳法を行っていたとしても原版と翻訳版との間での等価性の担保に懸念を生じさせる。

これらの尺度翻訳上の課題に対して，LLMの活用は，単なる翻訳作業の代替ではなく，むしろ意思決定過程を可視化し，検証可能性を高める補助的手段として有用な可能性がある。具体的には，LLMによって翻訳案を生成し，プロンプトや選択根拠を記録・共有することで，従来ブラックボックス化される傾向にあった翻訳判断に，より客観的な検証可能性と再現性を担保できる。例えば，*LLMTranslate* (Kunst, 2025) は，RからGPTに接続し，尺度の順翻訳，逆翻訳，比較を自動化するパッケージである。こうしたパッケージを使用することによって翻訳の負担軽減のみならず，プロンプトの公開による翻訳プロセスの透明性向上にも資すると期待される。実際，Kunst & Bierwiazonek (2023) は，国際的に広く使用されて

いるHEXACO性格検査を対象に，Google翻訳およびGPT-3.5と33言語の人手翻訳を比較し，人手翻訳に必ずしも劣らないことを示した。具体的には，当時のGPT-3.5よりGoogle翻訳の方が人手翻訳との類似性が高いものの，その差は小さかった。また，社会学者による7件法での評価においても，人手翻訳は最も高品質であると評価されたが，その効果量は小さかった。この結果から，同研究は，従来の翻訳手続きの代替になりうる，機械翻訳を活用した新たな翻訳フレームワークを提示し，多言語翻訳による実証研究の必要性を提案している。

AI翻訳の性能はテキストデータ量ではなく，原版言語と翻訳対象言語の系統的な距離（linguistic distance）に依存する。英語と日本語は系統的に大きく異なり，こうした言語距離は研究上の不利益につながる可能性が指摘されている (e.g., Cao et al., 2024)。Park & Oh (2025) は，中国語・韓国語・日本語において，人手翻訳とChatGPT-3.5，DeepLによる翻訳・逆翻訳を比較し，翻訳差を意味の一致度から検証した。その結果，翻訳精度の差について統計的検定は実施されていないが，質的比較によって，テキストデータ量が大きい中国語だけでなく，韓国語においても高い一致が観察された。一方，人手翻訳では，翻訳者の主観的解釈や言い換えにより，原文の意味から逸脱した場合がみられた。このような翻訳上の差異は，意味的等価性だけでなく，項目の解釈や回答者の心理的反応といった語用論的等価性にも影響しうる。したがって，言語的距離の大きい英語圏で開発された尺度を日本語に翻訳する際には，翻訳差が尺度特性に及ぼす影響を検証する必要がある。しかし，既存の研究では，尺度の信頼性，因子構造といった統計的特徴に，LLM等と人手といった翻訳の違いが及ぼす影響は検討されていない。

以上から，LLMを用いた尺度翻訳の有効性は期待できる一方，翻訳の違いによる尺度の等価性および統計的特徴への影響は明らかになっておらず，特に英語からの言語距離が大きい日本語においては検証が不可欠である。そこで本研究では，単一尺度に限定せず，異なる心理概念を測定する複数の尺度を対象に，LLM翻訳と人手翻訳による影響を検討することを目的とする。

ビッグチームサイエンス

本研究では，心理尺度の翻訳というこれまで個人や少人数の裁量に委ねられてきた作業とその過程を，ビッグチームサイエンス（Big Team Science: BTS）の方法を用いて大規模に検討する。BTS は，多数の研究者が知的・物的リソースを共有し，共通の理論的課題に対して協調的に取り組む枠組みであり，再現可能性と一般化可能性を高める新しい研究協働の仕組みとして注目されている (Coles, Hamlin et al., 2022; Forscher et al., 2023)。このアプローチは，中央集権的な大型プロジェクトとは異なり，研究者コミュニティが草の根的に組織され，各拠点が共通プロトコルのもとでデータを収集し，その成果を統合的に分析する点に特徴がある。心理学分野では，ManyLabs (Klein et al., 2014)，ManyBabies (Visser et al., 2022)，Psychological Science Accelerator (Moshontz et al., 2018) などがその代表例であり，これらの取り組みは多文化的環境における知見の再現性と透明性を大きく前進させてきた。このように多量の人的リソースを一気に注入できる方法は，既存知見における一

般化可能性の検証や追試などに適している (山田, 2024)。しかしながら，このような取り組みは日本国内の心理学領域では全く行われてこなかった。

多人数の研究者が協働するといっても，その形には様々なタイプが存在している。2010年代後半における「初期」の多著者論文では，主に意見論文を多数の著者が執筆する形式が多かった (e.g., Benjamin et al., 2017; Lakens et al., 2018; Trafimow et al., 2018)。これらは参加研究者らのリソースを持ち寄るというBTSの基盤とは少し異なっており，あくまで一定範囲でのコンセンサスの形成と署名を目的とした唐傘連判状のようなものであり，マルチシグナトリー型オピニオンと呼ぶことができる。最もオーソドックスなBTSの形態は，ある特定の仮説検証のために多くの研究者を呼び集め，それぞれの得意な仕事をサブチームとして割り当てるというフラッシュチーム型である (e.g., Coles, March et al., 2022; Parsons et al., 2022; Van Bavel et al., 2022)。このタイプの研究はデータ収集，翻訳，分析などが効率的にチーム単位で実行される利点を持っており，本研究でもこれに近い形態を取っている。もう一つは，多国間でのBTSにてよく見られるハブ・アンド・スポーク型である。この形態は，中心となるリードチームが各国のサブチームと個別にやり取りし，ターゲットとなる実験課題のローカライズを行う (e.g., Ruggeri et al., 2022)。サブチームごとにその内部の統治を委任できるため，リードチームの負担が比較的小さいことが特徴である。他にも，研究者を (少なくとも作業実施時には) 匿名のワーカーとして無作為に集める研究者クラウドソーシング型の研究も存在している (e.g., Breznau et al., 2022)。このように，BTSの型は研究対象や研究方法，研究時の情勢 (コロナ禍など) によってそれぞれ適したものがその場で選ばれるが，その方法論は依然として確立されておらず，試行錯誤が続いている。

本研究では，日本初の試みとして，このフラッシュチーム型のBTSのアプローチを翻訳研究に導入し，多人数の研究者が協働して多数の心理尺度を翻訳・評価する。これにより，日本国内の研究コミュニティでもBTSが可能であることを示し，かつ尺度翻訳におけるLLMの有用性を集中的，包括的に検証するという二重の目的を同時に達成しようとしている。

本プロジェクトの経緯

今回実施されたBTSの具体的なプロセスは以下の通りである。まず，本研究の発端となった議論は，SNS (X) で2025年9月8日に行われた一連のやりとりである (https://x.com/momentumyy/status/1964965876169617865)。ここでは主に*LLMTranslate*の科学的検証が必要であるという合意の形成がなされ，方法論としてはBTSがそれに適していることが提案された。そこで次に2025年9月12日，Discordにてメインリードオーサーらが研究の大枠をまとめ，本研究をManyScales Project (MSP) の主要課題として発足させた。

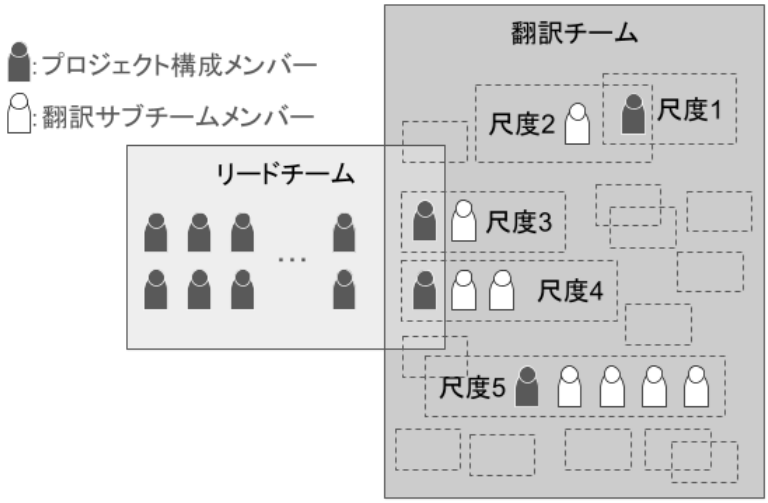
2025年9月26日，SNS (X) にて共同研究者の募集が開始された (https://x.com/asarin/status/1971348998905528692)。募集は一週間後の10月3日に締め切られたが，その間に計33名の応募があった。その中から共同研究契約書を締結した31名を第一次構成メンバーとして確定した。なお，共同研究契約時には各尺度翻訳作業にサブチームメンバーとして

参加予定の人物が連名している場合もあるが，ここではそのサブチーム代表者のみを本プロジェクトの構成メンバーとして数えている。

本プロジェクトの構成図を図1に示した。構成メンバーは，リードチーム（著者リストの最初の16名+最終著者）と翻訳チーム（それ以降の著者）に分けられた。メンバーによっては両チームに所属している場合もあり，その場合はリードチームのメンバーとして扱っている。リードチームは，*LLMTranslate*を使って尺度の翻訳，逆翻訳，および調整を行い，データを収集し，分析し，論文化する主体となるグループである。*LLMTranslate*を使うための費用（APIへの課金）とデータ収集にかかる費用もリードチームが負担する。翻訳チームのメンバーはそれぞれ対象にしたい英語の心理尺度を持ち寄った。そして，各尺度に対応する個人またはサブチーム単位で，それを*LLMTranslate*パッケージを使わないやり方で翻訳したものを提供することを求められた。

全ての情報共有とディスカッションはDiscord上で行われ，必要に応じて随時Zoom等によるオンラインでのミーティングが行われた。

図1. 本プロジェクトの構成



<u>著者順序の決定</u>

本プロジェクトでは共同研究者の募集時点で，上述した「リードチーム・翻訳チームに求められる貢献」が示され，参加希望者は自主的にいずれか，あるいは両方に参加することを選択した。

近年のBTSでは，著者順序をその貢献度に基づいて決定する"Tiered authorship model"が採用されることが多い (e.g., Abrams et al., 2025; Seminara et al., 2007)。これは，大規模な共同研究プロジェクトが一般化したことによって生じた，著者リストの長大化による著者責任の拡散に適切に対応すること，また謝辞（acknowledge）欄との境界のあいまい化

を解消し，論文への貢献を可視化する目的で誕生したモデルとされている (Cronin, 2001)。Tiered authorship modelでは，論文の構成および執筆・最終確認等，医学雑誌編集者国際委員会 (ICMJE) や各学会等が定めた実質的な著者貢献基準を満たす者を"Tier 1"，それ以外の限定的な貢献を果たした者を"Tier 2"などと呼称し[2]，論文の著者は (プロジェクトリーダー) Tier 1，Tier 2の順に何らかの順 (アルファベット順など) で表記されることが多い (e.g., DELVE, 2025; Vaidis et al., 2024; Van Bavel et al., 2022)。

これに倣い本論文の著者順序については，リードチームのメンバーが前半に配置され，翻訳を行う尺度の選定および原版著者への許可取りなどを部分的に担当した翻訳チームのメンバーが後に続くこと，およびチーム内での著者順は何らかの公平な手段によって決定される旨が共同研究者の募集時点で明言されており，プロジェクトメンバーはこれに同意した上でプロジェクトに参加している。

仮説設定

本研究では，リードチームを中心にDiscordやGoogleドキュメント上でのディスカッションと，翻訳チームも参加可能なZoom等によるオンラインでのミーティングにおけるディスカッションにより，検証する仮説を共同策定した[3]。LLM翻訳版と人手翻訳版の比較において，どちらが優れるのかどうかについては，リードチーム・翻訳チームに対してGoogleフォームを用いた調査を行って決定した。その結果，以下の4つの仮説が検証対象として採択された。

仮説１　LLM翻訳版と人手翻訳版に対して，心理尺度を研究で用いる心理学の専門家が，意味的忠実性，表現の自然さ，文化的妥当性に関する評価を行う。人手翻訳版の方が，LLM翻訳版よりも，意味的忠実性，文化的妥当性に関して有意に評価が高くなることが予想される。表現の自然さについては，人手翻訳版とLLM翻訳版との間に差は無いと予想される。

仮説２　LLM翻訳版と人手翻訳版を実際に人を対象とした調査を行った際，尺度の統計的特徴や受検者側の印象において違いがあるかどうかを検証する。

---

[2] Tiered authorship modelによる著者の序列化は，データや資料にどの程度アクセスできる環境にあるかといった点から地理的な影響を受ける可能性 (Adetula, 2022; Hoekman & Rake, 2024)，"Tier 2"と判断された若手研究者の貢献がキャリアパスにおいて無視されやすくなってしまう可能性など，いくつかの潜在的な問題を抱えている。以上の背景を踏まえてか，本論文で引用しているものを含め，近年のBTS論文の多くでは，本文中にTier制を示唆する表現を用いず，代わりにチームの具体的な役割に基づいた名称 (Leadership Team, Validation Teamなど) で表記されるケースが増えている。また近年では，著者一人ひとりの責任の範囲を明確化するために，CRediT (The Contributor Roles Taxonomy) の記載が要求される論文誌も増加しており，著者順のみに依存しないcontributorshipの評価の枠組みが広がっている (Holcombe, 2019)。そこで本研究でもTierという呼称は使用しないこととした。

[3] 心理尺度においては構成概念の妥当性も重要であることは論を俟たない。本研究においてもLLM翻訳版と人力翻訳版それぞれの妥当性についても検証を行うべきではある。ただ，文言の印象評定は表面的妥当性の検証にはなるが，構成概念妥当性の検証などは背景理論やその領域によって判じられるべきで，本研究のように複数の領域に跨った一般的な基準を作るのは難しい。また信頼性は妥当性の上限とも言われるように，まずは信頼性について多角的かつ一般的に検証を行うことが先決であり，本研究では妥当性に関して仮説を設定しない。

仮説2-1：尺度の構造的妥当性として尺度の背景理論（原論文）に基づく確認的因子分析を行い，モデル適合度に基づいて両者を比較した場合，LLM版も人手翻訳版もいずれも適合基準を満たす値が得られると予想される。

仮説2-2：尺度の実用的特徴として因子得点を算出し，LLM版と人手翻訳版を比較するが，平均点に統計的な差はみられないだろう。また尺度得点間の相関係数は0.75を超える高い相関が見られると予想される。

これらの仮説に加え，探索的に尺度を分析し，LLM版と人手翻訳版の記述的統計量の分析を行う。LLM版と人手尺度版のいずれにおいても，内的整合性の観点からはなんらかを測定するものであることは示されると予想される。これに加えて，探索的因子分析を行い，因子数や因子負荷量の値など，測定不変性の水準において同等かどうかを段階的に評価する。続いて確認的因子分析や尺度得点の算出を行い，その差異を検証する。

仮説３　一般回答者による評価では，文章の自然さや理解しやすさにおいて，LLM翻訳版と人手翻訳版とで統計的な差は見られないと予想される。

仮説４　LLM翻訳版と人手翻訳版に対して逆翻訳を行ったものと原版項目に対して，生成AIの埋め込み表現によるベクトル化を行い，コサイン類似度の評価を行う。人手翻訳版の方が，LLM翻訳版よりも，原版項目に対して有意にコサイン類似度が高いと予想される。

方法

翻訳

本研究では，心理尺度の翻訳精度と妥当性を比較するために，24の英語版尺度を対象とする。対象となった尺度は翻訳チームの研究者が，尺度の有用性や学術的価値などを元に選定したものを持ち寄ったものである。翻訳は二つの条件で行う。一つ目は LLM 翻訳条件であり，Rパッケージ *LLMTranslate* 0.2.0 (Kunst, 2025) を用いる。本パッケージはTRAPD (Translation, Review, Adjudication, Pretesting, Documentation) フレームワーク (Harkness, Villar, & Edward, 2010) やISPOR (International Society for Pharmacoeconomics and Outcomes Research) の国際ガイドライン (Wild et al., 2005) に準拠して設計されており，順翻訳と逆翻訳を自動的に実施できる機能を備えている。翻訳では*LLMTranslate*パッケージで使用可能（2025年11月29日現在）なOpenAI，Gemini，Claudeの各モデルで生成された訳語を比較検討し，いずれかの訳語を採用する。尺度ごとの順翻訳，逆翻訳，調整には異なるモデルが使用されうる。プロンプトは*LLMTranslate*デフォルトのものをすべての尺度で共通して使用する。出力の揺らぎについては，翻訳時の言い換えの幅を決定するパラメータであるTemperatureを，原文意味の保持および再現可能性の観点から0（言い換えの幅が最も小さい決定論的なもの）に設定する。必要に応じて複数回の生成を行い，その

中から代表訳を選択するが，この過程はすべて記録・公開し，再現可能性を確保する。また，*LLMTranslate*の機能の再現および拡張を目的として，本研究では*LLMTranslate*を基盤としつつ，複数の翻訳エンジンに対応可能なLLM翻訳用スクリプトを新たに作成した。作成したスクリプトは，以下の GitHubリポジトリ (https://github.com/VldmrSlvdr/ManyScales) にて公開している。

二つ目は人手翻訳条件であり，翻訳チームに属する複数の翻訳者によって日本語版尺度を作成する。翻訳者は辞書およびDeepLやGoogle翻訳を補助的に利用することは許されるが，生成AIとの対話に基づいた翻訳は認めない。翻訳の過程で使用したツールや手順，留意点については事前に報告させ，翻訳プロセスの透明性を担保する。

逆翻訳

作成された生成AIと人手翻訳版は，生成AIによる逆翻訳を行い，原版との対応を確認する。逆翻訳では生成AIによる翻訳に統一する理由としては，順翻訳時の各翻訳法の特徴を明らかにするのが目的のため，逆翻訳時の手法の違いが影響しないようにするためである。

翻訳プロセス

翻訳プロセスの設計はISPORの翻訳・文化適応ガイドライン (Wild et al., 2005) に準拠する。特に内容的妥当性を確保するため，LLM翻訳版と人手翻訳版の両方について，意味の等価性，概念の等価性，言語的自然さを中心に評価を行う。*LLMTranslate*による自動的な順翻訳と逆翻訳の手続きでは，原版との意味的一致を何らかの形で確認する必要がある。その上で，専門家による忠実性と自然さの評価，一般回答者による理解度と文化的妥当性の評価を組み合わせることで，ISPORが求める翻訳・文化適応の基準に沿った内容的妥当性の検証を実現する。

参加者

本研究には二種類の参加者を募集する。第一に，一般回答者として，日本語を母語とする成人を募集し，オンライン調査プラットフォーム (Qualtrics) にて回答を求める。予定サンプルサイズは4800名を目標とする。参加者には，翻訳条件の異なる尺度項目が提示される。提示順序はカウンターバランスを行い，順序効果の影響を統制する。

第二に，専門家評価者として，心理学研究の経験を持ち，英語と日本語に堪能な研究者や大学院生を募集する。各尺度につき少なくとも50名を確保する予定である。専門家には，人手と生成AIの翻訳文をブラインドかつランダムに提示し，各尺度に対する意味的忠実性，表現の自然さ，文化的妥当性の観点から評価を求める (5件法のリッカート)。

いずれの調査においても，オンライン調査におけるsatisficer (三浦・小林, 2015) とAIエージェントによる回答の影響をできるだけ緩和させるため，前者については注意チェック質問，後者については錯視などを用いた視覚的な認知トラップ質問 (Affonso, 2025) を挿入する。これら2問のいずれかに正答できなかった参加者は除外する。

サンプルサイズ設計と検出力分析

本研究のサンプルサイズは，心理測定学的比較に十分な検出力を確保することを目指して設定する。まず一般回答者サンプルについては，翻訳条件間の因子構造の比較を主たる目的とし，シミュレーション研究や既存の推奨 (e.g., Mundfrom et al., 2005) に基づき，尺度あたり少なくとも200名の回答者を確保することとした。したがって，24尺度を個別に実施する計画の下で，全体として4800名程度の参加者からデータを得ることを目標とする。

さらに専門家評価者サンプルについては，翻訳文ごとの忠実性や自然さのスコアを比較することを目的とする。内容的妥当性についてのサンプルサイズ設計については，COSMINの調査による内容的妥当性の基準 (Mokkink, Elsman, & Terwee, 2024) に従い，50名とする。

評価

尺度の評価は多面的に行う。心理測定学的分析として，内的一貫性係数の算出 (クロンバックの α係数等)，因子構造の確認 (平行分析)，測定不変性の確認 (確認的因子分析) を行う。回答者体験としては，理解度，自然さ，回答のしやすさ，翻訳が機械によるものかを識別できるかどうかを測定する。専門家評価については，翻訳条件間の差を統計的に比較する。さらに探索的分析として，埋め込みモデルを用いて翻訳文と英語原版をベクトル化し，項目ごとのベクトル間距離を算出する。指標としてはコサイン類似度やユークリッド距離を用い，多次元尺度構成法による可視化も行う。

使用尺度

本研究では，翻訳のために使用する対象として事前に翻訳チームメンバーが24種類の英語版尺度を選定した。詳細はTable 1に示している。

Table 1.
本研究で翻訳の対象とする原版心理尺度

	尺度名	概念	対象母集団	項目数	選択肢数	因子数	Index	信頼性	妥当性
1	Public Speaking Threats Inventory	人前で話すことの不安の理由	属性の限定なし (一般成人)	27	5	3	・各因子の項目の平均値 ・全項目の平均値	・α = .83 - .94 ・再検査ICC = .83 -.88	併存的妥当性 収束的妥当性
2	Revised Actively Open-minded Thinking About Evidence	積極的開放思考	属性の限定なし (一般成人)	8	6	1	・POMP得点 ・合計得点	・α = .71 - 74 ・再検査信頼性: 言及なし	予測的妥当性
3	Experiences in Human-AI Relationships Scale	AI愛着回避とAI愛着不安	生成AIの使用経験を有する成人	7	7	2	・2因子ごとの平均値	・α = .69, .79	内容的妥当性等
4	Expanded version of the Inventory of Depression and Anxiety Symptoms	感情障害の包括的な症状 (内在化)	属性の限定なし (一般成人) 患者	99	5	3	・各下位尺度 ・合計得点	・α = .72–.90 ・平均項目間相関 (AIC) = .34 - .74	弁別的妥当性 基準関連妥当性 構造的妥当性 収束的妥当性
5	Narcissistic Admiration and Rivalry Questionnaire	自己愛傾向 (Narcissism)	属性の限定なし (一般成人)	18	6	2	・平均値	・α = .73 - .88	法則定立 ネットワーク
6	Varieties of Sadistic Tendencies	サディズム (sadism)	属性の限定なし (一般成人)	16	5	2	・平均値	・α = .77 - .92	弁別的妥当性 併存的妥当性 構造的妥当性 収束妥当性
7	Moral Expansiveness Scale Short Form	道徳的な気遣いの拡張性	属性の限定なし (一般成人)	10	4	1	・平均値	・α = .84 ・再検査信頼性 r = .61, p < .001	予測的妥当性 弁別的妥当性 収束的妥当性 予測的妥当性
8	Self-control Strategy Scale	セルフコントロールの 方略使用の個人差	属性の限定なし (一般成人)	41	5	8	・下位尺度ごとの項目 回答値の平均値	・ω = .66 - 93	弁別的妥当性 収束的妥当性 予測的妥当性 増分妥当性
9	Orientation to Chocolate Questionnaire	チョコレートに関連する接近・回避 行動と罪悪感	属性の限定なし (一般成人)	14	9	3	・各因子の平均値	・α = .80 - .95	弁別的妥当性 併存的妥当性 構成概念妥当性
10	Tendency for Interpersonal Victimhood Scale	多様な対人関係場面にわたり 一貫して自己を"被害者"として認知 し続ける持続的感情	属性の限定なし (一般成人)	22	7	4	・各因子の平均値	・α = .85 - .90 ・再検査信頼性 r = .77	弁別的妥当性 構成概念妥当性 内容的妥当性 収束的妥当性 予測的妥当性
11	Open and Engaged State Questionnaire	心理的柔軟性	属性の限定なし (一般成人)	4	11	1	・合計得点	・α = .83 - .87	弁別的妥当性 収束的妥当性 増分妥当性
12	Psy-Flex	心理的柔軟性	属性の限定なし (一般成人)	6	5	1	・合計得点	・α = .78 - .97	弁別的妥当性 収束的妥当性 増分妥当性
13	PROMIS Sexual Function and Satisfaction Measures v2.0	性機能および満足度 尺度男性版	属性の限定なし (一般成人) 性機能障害患者	22	因子ごと 異なる	5	・合計得点	・α = .94 - .95	収束的妥当性
14	Fear of Being Single Scale (FOBS) Fear of Being Single in Relationships Scale (R-FOBS)	パートナー不在恐怖	属性の限定なし (一般成人) FOBS: 恋人・配偶者なし R-FOBS: 恋人・配偶者あり	6	5	1	・平均値	・α = .75 - .97 ・再検査信頼性: 言及なし	収束的妥当性 弁別的妥当性
15	Hedonic, Eudaimonic, and Extrinsic Motives for Activities (HEEMA) scale	動機づけ（志向性）としてのウェル・ビーイング	属性の限定なし (一般成人)	16	7	4	・平均値	・α = .79 - .91	言及なし 年代差のみ検証
16	Multidimensional Perfectionism Scale	多面的完全主義	属性の限定なし (一般成人)	45	7	3	・各下位尺度の合計	・α=.74 - .88	構成概念妥当性 収束的妥当性 弁別的妥当性 基準関連妥当性 併存的妥当性
17	Big Three Perfectionism Scale	完全主義 （多面的完全主義と異なる）	大学生	45	5	3	・各下位尺度の合計	・α = .87 - .95	因子的妥当性 収束的妥当性 弁別的妥当性 併存的妥当性
18	Usage Rating Profile-Intervention Revised	学校での介入の受容可能性 理解度，実現可能性，家庭連携，学校風土，システム支援	初等教育および 中等教育機関の教師	29	6	6	・合計得点	・α = .67 - .95	内容的妥当性 構成概念妥当性
19	Fear of Missing Out scale	取り残されることへの恐怖	属性の限定なし (一般成人)	10	5	1	・平均値	・α = .87	構成概念妥当性 収束的妥当性 弁別的妥当性 媒介的妥当性
20	Tech With Care Index for Teens	ポジティブなテクノロジー 利用の評価	13-17歳	17	5	**階層因子構造** **1次因子4** **2次因子2**	・各因子の平均値 ・総合TCI得点	・α = .60 - .80 ・再検査信頼性 r = .68 - .80	弁別的妥当性 収束的妥当性
21	Regulatory Mode Questionnaire	自己制御のモード	属性の限定なし (一般成人)	24	6	2	・平均値	・α = .57 - .85	構造的妥当性 構成概念妥当性 弁別的妥当性 収束的妥当性
22	Moral Outrage Scale	道徳的怒り（義憤）	属性の限定なし (一般成人)	9	5	2	・平均値	・α = .79 - .88	弁別的妥当性 収束的妥当性
23	Motivations to Eat Meat Inventory	肉食の動機づけ	属性の限定なし (一般成人)	19	7	4	・各因子の平均値 ・合計得点	・CFI > .98, TLI > .97, RMSEA < .05 ・α, 再検査信頼性: 記載なし	内容的妥当性 肉食者・菜食者間の測定不変性 収束体妥当性
24	General Attitudes Towards Artificial Intelligence Scale	AIに対する一般的な態度	属性の限定なし (一般成人)	20	5	2	・平均値	・α = .83, .88 ・再検査信頼性：言及なし	弁別的妥当性 構成概念妥当性 収束的妥当性

手続き

翻訳の生成と評価は，オンライン環境を通じて実施される。まず，各尺度についてLLM翻訳版と人手翻訳版を準備する。LLM翻訳には*LLMTranslate* 0.2.0を用い，共通プロンプトを設定して一括で実行する。人手翻訳は，翻訳者に各自の環境で実施してもらい，その後，提出された訳文を統一的なフォーマットに整理する。両条件で作成された日本語版は，自動または人手による逆翻訳を経て，原版との対応を確認する。

続いて，一般回答者にはQualtricsを介して尺度項目を提示する。翻訳条件はランダム化され，各参加者は一部の尺度に回答する。回答時間，回答のしやすさ，翻訳が機械的に生成されたものかどうかの認識についても5件法のリッカート尺度を用いて付加的に報告させる。専門家評価者には，両翻訳版の項目をブラインドかつランダムに提示し，尺度に対する意味的忠実性，表現の自然さ，文化的妥当性の観点から評価を求めるとともに (5件法のリッカート)，２つの翻訳のどちらが良いか強制選択を求める。

倫理

本研究のすべての手続きは，大阪大学大学院人間科学研究科行動学系研究倫理委員会の承認を得た上で実施する。参加者には事前に研究の概要を説明し，同意を得た上で回答を開始する。収集したデータは匿名化し，個人が特定されることがないように配慮する。また，本研究は心理学研究における国内 (日本心理学会倫理綱領) および国際 (AAPOR Code of Professional Ethics and Practices) 的な倫理規範に準拠して行う。本研究で扱うすべての尺度について，翻訳作業に先立って，翻訳者が原著者に対して個別に使用許諾を得ており，著作権や知的財産権に関する合意を遵守する形で進められる。

分析計画

LLM翻訳版と人手翻訳版の尺度それぞれに対して，以下の分析を行い両者の比較を試みる。

仮説１についての分析計画

専門家評価による意味的忠実性，表現の自然さ，文化的妥当性の評価に関して，LLM翻訳版と人手翻訳版とで母平均に差がないとするモデルと差があるとするモデルをベイズファクターで比較する。また，各尺度に対する各専門家の評定を個人レベル，集団平均を尺度レベルとし，さらに尺度レベルをまとめた上位階層をおいた階層モデルを考える。このとき最上位階層におけるLLM翻訳版評定平均と人手翻訳版評定平均の差の分布において，95%ベイズ信頼区間が0を含むかどうかを確認する。

仮説２についての分析計画

内的整合性の評価　LLM翻訳版と人手翻訳版双方の尺度において，クロンバックのα係数およびω係数を算出する。慣例に倣い，$\alpha > 0.8$であればいずれも全体としての一貫性が担保

されたものと判断する。ただし，元論文においてこの基準が満たされていない場合は，元論文のα係数を基準にする。

仮説2-1 尺度の構造的妥当性の検討　原論文に示された因子構造 (完全単純構造) を仮定した確認的因子分析を実行する。分析には*lavaan*パッケージ (Rosseel, 2012) を用いる。モデル適合度の評価には，χ^2/df (カイ二乗値をモデル自由度で除した値)，CFI (Comparative Fit Index)，SRMR (Standardized Root Mean Square Residual) を用いる。一般的な適合度の基準として，$\chi^2/df < 3$，CFI $\geq .95$，SRMR $\leq .08$を目安とする (Hu & Bentler, 1999)。同じ項目の翻訳間に等値制約を置いたモデルと置かないモデルを用意し，それぞれが適合度の判定基準を満たすかどうかを検証することで，構造不変，測定不変のどの水準まで一致しているかを検証できる。また，探索的な分析として，項目から得られた相関行列に対して平行分析 (Horn,1965) を適用し，共通因子数を求める。分析には*psych*パッケージ (Revelle, 2025) のfa.parallel関数を用い，因子数がLLM翻訳版と人手翻訳版とで一致するかどうかを判断する。

仮説2-2 尺度の実用的特徴　因子得点の算出は，原論文に得点算出の手順が明記されている場合はそれに従う。明記されていない場合は，完全単純構造を想定し，各因子に負荷する項目得点の和を因子得点とする。なお，探索的因子分析の結果が原論文の因子構造と異なる場合は，EFAに基づく因子構造でも追加的に得点を算出し，比較分析を行う。

LLM翻訳版と人手翻訳版それぞれについて，各因子の得点分布は詳細に検討する。具体的には，平均値，標準偏差，中央値，四分位範囲，最小値，最大値，歪度，尖度を算出する。これらの記述統計量の詳細は補足資料として提供する。

その上で，両版間の際を検討するため，対立する因子ごとに対応のある*t*検定を行う。ここでは比較にベイズファクターを使い，平均値に差がないモデルのほうが差があるモデルに比べて，相対的にデータに支持されると考える。さらに，LLM翻訳版と人手翻訳版の因子得点間のピアソン相関係数を算出する。これは個人レベルでの得点の対応関係を評価するものであり，$r \geq .90$であれば「両版の得点が個人レベルで高い一致を示す」と判断する。相関係数が高い場合，たとえ平均値に差があったとしても，個人の相対的順位は保たれていることを意味する。

これらの複数の指標を総合的に検討することで，LLM翻訳版と人手翻訳版の因子得点が実用上同等とみなせるかを多面的に評価する。

仮説３についての分析計画

一般回答者による文章の自然さや理解しやすさについての評価に関して，LLM翻訳版と人手翻訳版とで母平均に差がないとするモデルと差があるとするモデルをベイズファクターで比較する。また，各尺度に対する個々人の評定を個人レベル，集団平均を尺度レベルとし，さらに尺度レベルをまとめた上位階層をおいた階層モデルを考える。このとき最上位階層に

おけるLLM翻訳版評定平均と人手翻訳版評定平均の差の分布において，95%ベイズ信頼区間が0を含むかどうかを確認する。

仮説４についての分析計画

　LLM翻訳版と人手翻訳版の逆翻訳と原版項目とのコサイン類似度を算出し，比較を行う。具体的には，各尺度について，LLM翻訳版および人手翻訳版の逆翻訳テキストと原版項目テキストを，多言語対応の言語モデル (HuggingFaceでの公開モデル，paraphrase-multilingual-MiniLM-L12-v2やLaBSEなど) を用いて埋め込み，ベクトル表現を取得する。それぞれの埋め込みベクトルについて，原版と LLM翻訳版の逆翻訳，原版と人手翻訳版の逆翻訳のベクトルペアごとにコサイン類似度を算出する。得られた類似度指標について，適切な統計的手法を用いて両翻訳の差異を比較する。

探索的検討

　本研究では，仮説検証に加えていくつかの探索的分析も実施する予定である。まず，プロンプトの内容が翻訳精度に及ぼす影響について検討する。特に，尺度作成において研究者が重視する観点 (概念的忠実性や回答者への配慮など) をプロンプトに含めた場合と含めない場合で，翻訳の心理測定学的特性や回答者評価がどのように変化するかを明らかにする。また，LLM翻訳における出力の確率的揺らぎを定量化し，単回の翻訳で十分とみなせるか，それとも複数出力の中から最良の訳を選定すべきかを検討する。さらに，文化適応が求められる語や項目 (例えば「party」) については，LLM翻訳と人手翻訳の扱い方を比較することで，文化的要因が翻訳の質や回答のしやすさに及ぼす影響を探索的に分析する。加えて，将来的な応用を視野に入れ，LLM翻訳によって年齢層や教育水準に応じた表現調整 (例えば小学生向けの簡易化) が可能かどうか，ならびにその妥当性を評価する。これらの探索的検討は，翻訳研究における新たな可能性を示すものであり，主要仮説の検証を補完する位置づけを持つ。

オープンプラクティス宣言

　解析に用いたRやPythonコードはGithubリポジトリ (https://github.com/VldmrSlvdr/ManyScales) にて公開する。プレレジ，データ，補足資料は全てOSF (https://osf.io/hn4fr/) にて公開する。一方で，本研究では翻訳した尺度については，それぞれの検証は行っておらず，正式な日本語版としての公表を目的としていないため，公開しない。

引用文献

Abrams, E., Leone, P. V., Cambrosio, A., & Faraj, S. (2025). The governance of open science: A comparative analysis of two open science consortia. *Research Policy*, *54*(3), Article 105195. https://doi.org/10.1016/j.respol.2025.105195

Adetula, A., Forscher, P. S., Basnight-Brown, D., Azouaghe, S., & IJzerman, H. (2022). Psychology should generalize from—not just to—Africa. *Nature Reviews Psychology*, *1*(7), 370–371. https://doi.org/10.1038/s44159-022-00070-y

Affonso, F. M. (2025). Detecting vision-enabled AI respondents in behavioral research through cognitive traps. *PsyArXiv*. https://doi.org/10.31234/osf.io/enuqj_v1

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, *183*(6), 589–596. https://doi.org/10.1001/jamainternmed.2023.1838

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., Griffiths, T. L., Haridi, S., Jagadish, A. K., Ji-An, L., Kipnis, A., Kumar, S., Ludwig, T., Mathony, M., Mattar, M., … Schulz, E. (2025). A foundation model to predict and capture human cognition. *Nature*, *644*(8078), 1002–1009. https://doi.org/10.1038/s41586-025-09215-4

Binz, M., Alaniz, S., Roskies, A., Aczel, B., Bergstrom, C. T., Allen, C., Schad, D., Wulff, D., West, J. D., Zhang, Q., Shiffrin, R. M., Gershman, S. J., Popov, V., Bender, E. M., Marelli, M., Botvinick, M. M., Akata, Z., & Schulz, E. (2025). How should the advancement of large language models affect the practice of science? *Proceedings of the National Academy of Sciences of the United States of America*, *122*(5), e2401227121. https://doi.org/10.1073/pnas.2401227121

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

Bowers, J. S., Puebla, G., Thorat, S., Tsetsos, K., & Ludwig, C. J. H. (2025). Centaur: A model without a theory. *PsyArXiv*. https://doi.org/10.31234/osf.io/v9w37_v3

Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H. V., Adem, M., Adriaans, J., Alvarez-Benjumea, A., Andersen, H. K., Auer, D., Azevedo, F., Bahnsen, O., Balzer, D., Bauer, G., Bauer, P. C., Baumann, M., Baute, S., Benoit, V., Bernauer, J., Berning, C., … Żółtak, T. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(44), e2203150119. https://doi.org/10.1073/pnas.2203150119

Brislin R. W. (1986). The wording and translation of research instruments. In Lonner W., Berry J. (Eds.), *Field Methods in Cross-Cultural Research* (pp. 137–164). Sage.

Cao, Y., Sickles, R. C., Triebs, T. P., & Tumlinson, J. (2024). Linguistic distance to English impedes research performance. *Research Policy*, *53*(4), 104971. https://doi.org/10.1016/j.respol.2024.104971

Coles, N. A., Hamlin, J. K., Sullivan, L. L., Parker, T. H., & Altschul, D. (2022). Build up big-team science. *Nature*, *601*(7894), 505–507. https://doi.org/10.1038/d41586-022-00150-2

Coles, N. A., March, D. S., Marmolejo-Ramos, F., Larsen, J. T., Arinze, N. C., Ndukaihe, I. L. G., Willis, M. L., Foroni, F., Reggev, N., Mokady, A., Forscher, P. S., Hunter, J. F., Kaminski, G., Yüvrük, E., Kapucu, A., Nagy, T., Hajdu, N., Tejada, J., Freitag, R. M. K., … Liuzza, M. T. (2022). A multi-lab test of the facial feedback hypothesis by the Many Smiles Collaboration. *Nature Human Behaviour*, *6*(12), 1731–1742. https://doi.org/10.1038/s41562-022-01458-9

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology*, *52*(7), 558–569. https://doi.org/10.1002/asi.1097

Cross, J., Kayalackakom, T., Robinson, R. E., Vaughans, A., Sebastian, R., Hood, R., Lewis, C., Devaraju, S., Honnavar, P., Naik, S., Joseph, J., Anand, N., Mohammed, A., Johnson, A., Cohen, E., Adeniji, T., Nnenna Nnaji, A., & George, J. E. (2025). Assessing ChatGPT's capability as a new age standardized patient: Qualitative study. *JMIR Medical Education*, *11*(1), e63353. https://doi.org/10.2196/63353

Cruchinho, P., López-Franco, M. D., Capelas, M. L., Almeida, S., Bennett, P. M., Miranda da Silva, M., Teixeira, G., Nunes, E., Lucas, P., Gaspar, F., & Handovers4Safe Care. (2024). Translation, cross-cultural adaptation, and validation of measure

ment instruments: A practical guideline for novice researchers. *Journal of Multidisciplinary Healthcare*, *17*, 2701–2728. https://doi.org/10.2147/JMDH.S419714

DELVE (DECam Local Volume Exploration) Survey. (2025). DELVE Policy Guidelines Version 2.3. https://delve-survey.github.io/docs/DELVE_PolicyGuidelines.pdf

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Elyoseph, Z., Hadar-Shoval, D., Asraf, K., & Lvovsky, M. (2023). ChatGPT outperforms humans in emotional awareness evaluations. *Frontiers in Psychology*, *14*, 1199058. https://doi.org/10.3389/fpsyg.2023.1199058

Epstein, J., Santo, R. M., & Guillemin, F. (2015). A review of guidelines for cross-cultural adaptation of questionnaires could not bring out a consensus. *Journal of Clinical Epidemiology*, 68, 435-441. http://dx.doi.org/10.1016/j.jclinepi.2014.11.021

Forscher, P. S., Wagenmakers, E.-J., Coles, N. A., Silan, M. A., Dutra, N., Basnight-Brown, D., & IJzerman, H. (2023). The benefits, barriers, and risks of big-team science. *Perspectives on Psychological Science*, *18*(3), 607–623. https://doi.org/10.1177/17456916221082970

Friese, S. P. (2025). Conversational analysis with AI - CA to the power of AI: Rethinking coding in qualitative analysis. *OSF Preprints*. https://doi.org/10.31219/osf.io/6b52m_v1

Google. (2025). Gmail's protections are strong and effective, and claims of a major Gmail security warning are false. The Keyword. https://blog.google/products/workspace/gmail-security-protections/

Götz, F. M., Maertens, R., Loomba, S., & van der Linden, S. (2024). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods*, *29*(3), 494–518. https://doi.org/10.1037/met0000540

Granas, A. G., Nørgaard, L. S., & Sporrong, S. K. (2014). Lost in translation?: Comparing three Scandinavian translations of the Beliefs about Medicines Questionnaire. *Patient Education and Counseling*, *96*(2), 216–221. https://doi.org/10.1016/j.pec.2014.05.010

Harkness J. (2003). Questionnaire translation. In Harkness J. A., van de Vijver F. J. R., Mohler P. P. (Eds.), *Cross-Cultural Survey Methods* (pp. 35–56). Wiley.

Harkness J. A., Villar A., & Edwards B. (2010). Translation, adaptation, and design. In Harkness J. A.et al. (Eds.), *Survey Methods in Multinational, Multicultural and Multiregional Contexts* (pp. 117-140). Hoboken, NJ: John Wiley.

Heinz, M. V., Mackin, D. M., Trudeau, B. M., Bhattacharya, S., Wang, Y., Banta, H. A., J ewett, A. D., Salzhauer, A. J., Griffin, T. Z., & Jacobson, N. C. (2025). Randomiz ed trial of a generative AI chatbot for mental health treatment. *NEJM AI*, *2*(4). https://doi.org/10.1056/aioa2400802

Hoekman, J., & Rake, B. (2024). Geography of authorship: How geography shapes a uthorship attribution in big team science. *Research Policy*, *53*(2), 104927. https://doi.org/10.1016/j.respol.2023.104927

Holcombe, A. (2019). Farewell authors, hello contributors. *Nature*, *571*(7764), 147. https://doi.org/10.1038/d41586-019-02084-8

Horn J.L. (1965). A rationale and test for the number of factors in factor analysis. *P sychometrika, 30*(2), 179–185. https://doi.org/10.1007/BF02289447

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Mo deling: A Multidisciplinary Journal, 6*(1), 1-55. https://doi.org/10.1080/10705 519909540118

Huang, M., Zhang, X., Soto, C., & Evans, J. (2024). Designing LLM-agents with perso nalities: A psychometric approach. *arXiv*. https://doi.org/10.48550/arXiv.241 0.19238

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocia n, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., C heong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Gallia ni, E. M., ⋯ Nosek, B. A. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology*, *45*(3), 142–152. https://doi.org/1 0.1027/1864-9335/a000178

Kunst, J. R. (2025). *LLMTranslate: 'shiny' app for TRAPD/ISPOR survey translation w ith LLMs* (R package version 0.1.3). Comprehensive R Archive Network (CRA N). https://doi.org/10.32614/CRAN.package.LLMTranslate

Kunst, J. R., & Bierwiaczonek, K. (2023). Utilizing AI questionnaire translations in cro ss-cultural and intercultural research: Insights and recommendations. *Internat ional Journal of Intercultural Relations: IJIR*, *97*(101888), 101888. https://doi. org/10.1016/j.ijintrel.2023.101888

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Bagul ey, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins,

G. S., Crook, Z., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x

Liu, W., & Ding, N. (2025). Can Centaur truly simulate human cognition? The fundamental limitation of instruction understanding. PsyArXiv. https://doi.org/10.31234/osf.io/zfhv9_v1

Maertens, R., Götz, F. M., Golino, H. F., Roozenbeek, J., Schneider, C. R., Kyrychenko, Y., Kerr, J. R., Stieger, S., McClanahan, W. P., Drabot, K., He, J., & van der Linden, S. (2024). The Misinformation Susceptibility Test (MIST): A psychometrically validated measure of news veracity discernment. *Behavior Research Methods*, *56*(3), 1863–1899. https://doi.org/10.3758/s13428-023-02124-2

Massé, C. C., Krieger, V., Peró-Cebollero, M., Amador-Campos, J. A., & Guàrdia-Olmos, J. (2025). Measurement invariance and cross-linguistic validation of the PSS-4 in university context: multidimensional analysis and associations with psychological and behavioral outcomes. *Frontiers in Psychology*, *16*(1648070), 1648070. https://doi.org/10.3389/fpsyg.2025.1648070

Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., Capstick, E., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., Walsh, T., Hamrah, A., Santarlasci, L., … Oak, S. (2025). Artificial Intelligence Index Report 2025. *arXiv*. https://doi.org/10.48550/arXiv.2504.07139

Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(9), e2313925121. https://doi.org/10.1073/pnas.2313925121

三浦 麻子・小林 哲郎 (2015). オンライン調査モニタのSatisficeに関する実験的研究. 社会心理学研究, *31*(1), 1–12. https://doi.org/10.14966/jssp.31.1_1

Mokkink, L. B., Elsman, E. B. M., & Terwee, C. B. (2024). COSMIN guideline for systematic reviews of patient-reported outcome measures version 2.0. *Quality of Life Research, 33*(11), 2929–2939. https://doi.org/10.1007/s11136-024-03761-6

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., Antfolk, J., Castille, C. M., Evans, T. R., Fiedler, S., Flake, J. K., Forero, D. A., Janssen, S. M. J., Keene, J. R., Protzko, J., Aczel, B., … Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501–515. https://doi.org/10.1177/2515245918797607

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, *5*(2), 159–168. https://doi.org/10.1207/s15327574ijt0502_4

Namazova, S., Brondetta, A., Strittmatter, Y., Nassar, M., & Musslick, S. (2025). Not yet AlphaFold for the mind: Evaluating Centaur as a synthetic participant. *arXiv*. https://doi.org/10.48550/arXiv.2508.07887

Orr, M., Cranford, D., Ford, K., Gluck, K., Hancock, W., Lebiere, C., Pirolli, P., Ritter, F., & Stocco, A. (2025). Not even wrong: On the limits of prediction as explanation in cognitive science. *arXiv*. http://arxiv.org/abs/2510.03311

Ozolins, U., Hale, S., Cheng, X., Hyatt, A., & Schofield, P. (2020). Translation and back-translation methodology in health research – a critique. *Expert Review of Pharmacoeconomics & Outcomes Research*, 1–9. https://doi.org/10.1080/14737167.2020.1734453

Park, J. J., & Oh, J. (2025). Enhancing international research through alternative back translation methods leveraging artificial intelligence. *Human Resource Development International*, 1–22. https://doi.org/10.1080/13678868.2025.2558571

Parsons, S., Azevedo, F., Elsherif, M. M., Guay, S., Shahim, O. N., Govaart, G. H., Norris, E., O'Mahony, A., Parker, A. J., Todorovic, A., Pennington, C. R., Garcia-Pelegrin, E., Lazić, A., Robertson, O., Middleton, S. L., Valentini, B., McCuaig, J., Baker, B. J., Collins, E., … Aczel, B. (2022). A community-sourced glossary of open scholarship terms. *Nature Human Behaviour*, *6*(3), 312–318. https://doi.org/10.1038/s41562-021-01269-4

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Ruggeri, K., Panin, A., Vdovic, M., Većkalov, B., Abdul-Salaam, N., Achterberg, J., Akil, C., Amatya, J., Amatya, K., Andersen, T. L., Aquino, S. D., Arunasalam, A., Ashcroft-Jones, S., Askelund, A. D., Ayacaxli, N., Sheshdeh, A. B., Bailey, A., Barea Arroyo, P., Mejía, G. B., … García-Garzon, E. (2022). The globalizability of temporal discounting. *Nature Human Behaviour*, *6*(10), 1386–1397. https://doi.org/10.1038/s41562-022-01392-w

Salama-Younes, M., Montazeri, A., Ismaïl, A., & Roncin, C. (2009). Factor structure and internal consistency of the 12-item General Health Questionnaire (GHQ-12) and the Subjective Vitality Scale (VS), and the relationship between them: a study from France. *Health and Quality of Life Outcomes*, *7*(1), 22. https://doi.org/10.1186/1477-7525-7-22

Sanz, A., Tapia, J. L., García-Carpintero, E., Rocabado, J. F., & Pedrajas, L. M. (2025). ChatGPT simulated patient: Use in clinical training in psychology. *Psicothema*, *37*(3), 23–32. https://doi.org/10.70478/psicothema.2025.37.21

佐々木 研一・豊田 秀樹 (2024). ChatGPTにより生成された心理尺度項目の信頼性・妥当性の評価　日本テスト学会誌, *20*(1), 111–133. https://doi.org/10.24690/jart.20.1_111

Schröder, S., Morgenroth, T., Kuhl, U., Vaquet, V., & Paaßen, B. (2025). Large Language Models do not simulate human psychology. *arXiv*. https://doi.org/10.48550/arXiv.2508.06950

Seminara, D., Khoury, M. J., O'Brien, T. R., Manolio, T., Gwinn, M. L., Little, J., Higgins, J. P. T., Bernstein, J. L., Boffetta, P., Bondy, M., Bray, M. S., Brenchley, P. E., Buffler, P. A., Casas, J. P., Chokkalingam, A. P., Danesh, J., Smith, G. D., Dolan, S., Duncan, R., … Ioannidis, J. P. A. (2007). The emergence of networks in human genome epidemiology: Challenges and opportunities. *Epidemiology*, *18*(1), 1–8. https://doi.org/10.1097/01.ede.0000249540.17855.b7

Stefana, A., Damiani, S., Granziol, U., Provenzani, U., Solmi, M., Youngstrom, E. A., & Fusar-Poli, P. (2024). Psychological, psychiatric, and behavioral sciences measurement scales: best practice guidelines for their development and validation. *Frontiers in Psychology*, 15, 1494261. https://doi.org/10.3389/fpsyg.2024.1494261

Symeonaki, M., Stamou, G., Kazani, A., Tsouparopoulou, E., & Stamatopoulou, G. (2024). Examining the development of attitude scales using Large Language Models (LLMs). *arXiv preprint*. https://doi.org/10.48550/arXiv.2405.19011

Teixeira da Silva, J. A., & Yamada, Y. (2024). Could generative artificial intelligence serve as a psychological counselor? Prospects and limitations. *Central Asian Journal of Medical Hypotheses and Ethics*, *5*(4), 297–303. https://doi.org/10.47316/cajmhe.2024.5.4.06

Trafimow, D., Amrhein, V., Areshenkoff, C. N., Barrera-Causil, C. J., Beh, E. J., Bilgiç, Y. K., Bono, R., Bradley, M. T., Briggs, W. M., Cepeda-Freyre, H. A., Chaigneau, S. E., Ciocca, D. R., Correa, J. C., Cousineau, D., de Boer, M. R., Dhar, S. S., Dolgov, I., Gómez-Benito, J., Grendar, M., … Marmolejo-Ramos, F. (2018). Manipulating the alpha level cannot cure significance testing. *Frontiers in Psychology*, *9*, 699. https://doi.org/10.3389/fpsyg.2018.00699

Vaidis, D. C., Sleegers, W. W. A., van Leeuwen, F., DeMarree, K. G., Sætrevik, B., Ross, R. M., Schmidt, K., Protzko, J., Morvinski, C., Ghasemi, O., Roberts, A. J., Stone, J., Bran, A., Gourdon-Kanhukamwe, A., Gunsoy, C., Moussaoui, L. S., Smith, A. R., Nugier, A., Fayant, M.-P., … Priolo, D. (2024). A multilab replication of th

e induced-compliance paradigm of cognitive dissonance. *Advances in Methods and Practices in Psychological Science*, *7*(1), 25152459231213375. https://doi.org/10.1177/25152459231213375

Van Bavel, J. J., Cichocka, A., Capraro, V., Sjåstad, H., Nezlek, J. B., Pavlović, T., Alfano, M., Gelfand, M. J., Azevedo, F., Birtel, M. D., Cislak, A., Lockwood, P. L., Ross, R. M., Abts, K., Agadullina, E., Aruta, J. J. B., Besharati, S. N., Bor, A., Choma, B. L., … Boggio, P. S. (2022). National identity predicts public health support during a global pandemic. *Nature Communications*, *13*, 517. https://doi.org/10.1038/s41467-021-27668-9

Visser, I., Bergmann, C., Byers-Heinlein, K., Dal Ben, R., Duch, W., Forbes, S., Franchin, L., Frank, M. C., Geraci, A., Hamlin, J. K., Kaldy, Z., Kulke, L., Laverty, C., Lew-Williams, C., Mateu, V., Mayor, J., Moreau, D., Nomikou, I., Schuwerk, T., … Zettersten, M. (2022). Improving the generalizability of infant psychological research: The ManyBabies model. *Behavioral and Brain Sciences*, *45*(e35), e35. https://doi.org/10.1017/S0140525X21000455

Werner, P., Eliyahu, E., & Krupat, E. (2025). Mapping the translation and psychometric characteristics of the Patient-Practitioner Oriented Scale: A scoping review. *Patient Education and Counseling*, *137*(108787), 108787. https://doi.org/10.1016/j.pec.2025.108787

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., Erikson, P., & ISPOR Task Force for Translation and Cultural Adaptation (2005). Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: Report of the ISPOR task force for translation and cultural adaptation. *Value in Health, 8*(2), 94–104. https://doi.org/10.1111/j.1524-4733.2005.04054.x

Revelle, W. (2025). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 2.5.6, https://CRAN.R-project.org/package=psych.

山田 祐樹 (2024). 心理学を遊撃する――再現性問題は恥だが役に立つ――. ちとせプレス.

Zhang, W., Balloo, K., Hosein, A., & Medland, E. (2024). A scoping review of well-being measures: Conceptualisation and scales for overall well-being. *BMC Psychology*, *12*(1), 585. https://doi.org/10.1186/s40359-024-02074-0

著者のORCID情報

山田 祐樹　0000-0003-1431-568X
小杉 考司　0000-0001-5816-0099
国里 愛彦　0000-0002-5830-7182
分寺 杏介　0000-0002-9512-7439
後藤 崇志　0000-0002-6068-2897
橋本 泰央　0000-0002-5585-5247
工藤 大介　0009-0009-3720-0501
李 禕飛　0000-0001-6215-3238
眞嶋 良全　0000-0003-3579-1546
向井 智哉　0000-0002-4237-6781
野村 竜也　0000-0002-4391-9792
小口 真奈　0000-0002-1042-198X
七條 花恋　0009-0006-2466-345X
下司 忠大　0000-0001-6917-7945
高松 礼奈　0000-0002-9044-5446
竹橋 洋毅　なし
竹下 昌志　0000-0001-5853-3262
浅野 良輔　0000-0001-5812-5859
福田 実奈　0000-0002-6637-4270
古谷 嘉一郎　0000-0002-6370-4975
日道 俊之　0000-0003-2073-854X
平野 寛樹　0000-0001-9353-0147
五十嵐 祐　0000-0001-9432-4425
伊藤 雅隆　0000-0002-2178-0734
香川 璃奈　0000-0002-0482-5179
神野 雄　なし
加藤 弘通　なし
古村 健太郎　0000-0002-4720-3265
宮川 裕基　0000-0003-0529-3124
水野 君平　0000-0002-9308-2759
村浦 新之助　0009-0006-7125-1984
新谷 優　0000-0002-2733-3724
西村 多久磨　0000-0002-3508-8634
尾崎 由佳　0000-0001-8896-1081
佐藤 秀樹　0000-0001-5057-078X
佐藤 奈月　0009-0001-7710-9456

嶋 大樹　0000-0002-9566-2888
瀧川 諒子　0000-0002-2246-5664
田中 勝則　0000-0001-9650-8316
塚本 早織　0000-0001-6402-3867
山崎 茜　0000-0002-5093-6785
楊 帆　0000-0002-7961-3111
三浦 麻子　0000-0002-7563-7503