



Discussion Papers In Economics And Business

日本におけるワーク・エンゲイジメントスコアのレスポンス・スタイル補正

—係留ビネット法・多次元 IRT モデルを用いて—

澤田 真央

開本 浩矢

Discussion Paper 26-12

September 2026

Graduate School of Economics
The University of Osaka, Toyonaka, Osaka 560-0043, JAPAN

日本におけるワーク・エンゲイジメントスコアのレスポンス・スタイル補正—係留ビネット
法・多次元 IRT モデルを用いて—

澤田 真央[†]

開本 浩矢[‡]

要旨

本研究の目的は、日本国内で就業する従業員のワーク・エンゲイジメント（WE）スコアが国際的に低水準であることに對し、先行研究で指摘されている文化的なバイアスが影響している可能性の検証として、回答時のバイアスであるレスポンス・スタイル（RS）に着目し、どの程度スコアに影響しているかを定量的に検討することである。インターネット調査を通じて正規雇用者 442 名を対象に、超短縮版 UWES の 3 項目と 6 つのビネット項目を含む質問紙調査を実施した。分析には、係留ビネット法と多次元項目反応理論を組み合わせたベイズ統計解析を用い、回答者の RS を統制したモデルを推定した。その結果、モデル比較の観点では RS の影響の存在が示唆された一方で、ほぼすべての回答者において意味のある大きさを持つと判断できる RS は認められず、WE スコアに実質的な影響を与えているとは言えなかった。また、RS を統制した場合の補正効果は 0.013 ポイントと極めて小さく、先行研究で報告されている国際比較差（約 1 ポイント）の 1.3%に相当するにすぎず、効果量（Cohen's $d = 0.010$ ）も実質的に意味のない水準であった。以上の結果から、日本人の WE の低さは RS では説明することが難しく、項目内容に依存するような他のバイアスの影響や、職務特性や職場環境など他の要因を検討する必要があることが示唆された。

JEL 分類：O15 J24 J28

キーワード：ワーク・エンゲイジメント、ユトレヒト・ワーク・エンゲイジメント尺度、文化バイアス、係留ビネット法、多次元 IRT モデル

[†] 大阪大学経済学研究科 経営学系専攻 博士後期課程 mail: u338546k@ecs.osaka-u.ac.jp

[‡] 大阪大学経済学研究科, mail: hirakimoto@econ.osaka-u.ac.jp

1.はじめに

本研究の目的は、先行研究で指摘されてきた日本人労働者のワーク・エンゲイジメント（Work Engagement, 以下 WE）の国際的な低水準に対して、回答時のバイアスがどの程度寄与しているのかを、係留ビネット法（Anchoring Vignettes）と多次元 IRT (Item Response Theory, 以下 IRT) モデルを用いて定量的に検討することである。これまでの先行研究では、WE の国際比較研究は蓄積されており、日本人の WE スコアが他国と比べて低いことが報告されている（Shimazu et al., 2010; 島津, 2014）。この測定上の問題の原因として、Shimazu et al. (2010) は文化的なバイアスの影響を指摘している。日本では集団の調和を重視する文化的背景から、ポジティブな感情や態度の表出を抑制する傾向があることを指摘しており（Iwata et al., 1995）、こうした文化的傾向が WE スコアの低水準の一因となっている可能性を示唆している。このような回答時のバイアスには代表的なものとして、レスポンス・スタイル（Response Style, 以下 RS）が挙げられる。RS とは質問の内容に関係なく、特定の選択肢を選択する傾向と定義され（Johnson et al., 2011; Yang et al., 2010）、日本人には極端な選択肢を避け中間的な選択肢を選びやすいといった文化的な回答傾向があることが知られている（Chen et al., 1995; 米村・松田, 1991）。RS は心理尺度を用いた測定において信頼性及び妥当性を脅かす要因として議論されており（Chen et al., 1995）、WE の測定尺度である UWES（Utrecht Work Engagement Scale, 以下 UWES）においても同様の影響が生じる可能性がある。しかしながら、RS が WE スコアをどの程度歪めているのか、またその影響が国際比較で観察される日本人の低スコアをどの程度説明しうるのかを定量的に検証した研究はこれまで存在しない。こうした問題意識のもと、本研究では、(1) 日本の従業員の WE スコアに RS が実質的な影響を与えているのか、(2) RS による WE スコアへの集団レベルの効果は、先行研究の国際比較における日本と他国のスコア差を説明するのかを検証し、日本の WE スコアへの RS による影響を考察する。

2. 先行研究と理論的背景

2.1 WE の定義と測定尺度

WE は、Kahn (1990) のパーソナル・エンゲイジメントの概念枠組みに基づき Schaufeli et al. (2002) によって概念化された。WE は、特定の出来事や対象に一時的に生じる感情ではなく、仕事に向けられた持続的かつ全般的な感情と認知であり、活力（vigor）、熱意（dedication）、没頭（absorption）という三つの側面によって特徴づけられる（Schaufeli et al., 2002; Schaufeli & Bakker, 2004）。本研究でも、多くの先行研究と同様に、この定義に従うものとする。WE を測定する代表的な尺度である UWES は、三つの下位次元から構成され、17 項目で活力、熱意、没頭を測定する（Schaufeli et al., 2002）。これまでに、32 カ国で標準化または使用されており、日本語版を含む、いずれの言語においても、良好な信頼性及び妥当性が確認されている（島津,

2025; Shimazu et al., 2008)。

2.2 WE スコアの国際比較における日本人の水準

WE スコアを国際比較した研究では、日本の WE スコアが著しく低い値となっていることが示されている。Shimazu et al. (2010) は、UWES の短縮版 (9 項目版) を用いて日本を含む 16 カ国において WE スコア比較しており、そのなかで日本のスコア平均は 2 点台後半と最下位であったのに対し、他の 15 カ国は 3 点台後半を上回っており、日本の次に低い値であるスペインでは 3 点台後半、最上位のフランスでは 4 点台後半となっていた。

さらに、Shimazu et al. (2010) は日本とオランダの WE スコアのデータに対し IRT を適用した分析では、日本のサンプルにおける UWES 尺度の測定精度が低い可能性を報告しており、日本人の WE スコアが極端に低い要因として日本人がポジティブな感情を表出するのを抑制する傾向 (Iwata et al., 1995) が関連していると結論づけている。これらの先行研究を踏まえると、WE スコアの正確な測定や、国際比較における日本人の WE スコアの議論には、バイアスの影響の定量的な測定・評価や、補正の手法が重要になる。

2.3 RS と日本人の特徴

RS は質問の内容に関係なく、特定の選択肢を選択する傾向と定義される (Johnson et al., 2011; Yang et al., 2010)。大きく、黙従反応傾向 (acquiescence response style)、極端反応傾向 (extreme response style)、中間反応傾向 (midpoint response style) の 3 つに分類される (Yang et al., 2010)。RS は、平均値や相関係数、回帰係数などの推定値が歪められる可能性があるため、調査結果の妥当性に対する脅威となる (Van Vaerenbergh & Thomas, 2013)。さらに、RS は文化的・社会的要因と関連しており、国際調査においては、比較可能性を阻む要因となる (Kankaraš & Moors, 2011)。なかでも、日本人の RS に関しては、極端な回答を避け、中間的な回答を好む傾向にあることなどが複数の先行研究において報告されている (Chen et al., 1995; 米村・松田, 1991; 田崎・申, 2017)。これらの知見を踏まえると、特に UWES はポジティブ項目のみで構成されるため、日本人の RS が WE スコアを低下させている可能性が示唆される。

2.4 係留ビネット法と IRT によるバイアス補正

リッカート尺度のような順序尺度を用いた調査において RS を統制し、尺度得点の比較可能性を高める方法として、係留ビネット法が提案されている。通常の自己評定の質問項目では回答者の潜在特性と RS の両方がスコアに反映されるため、自己評定データのみからこれらを分離することは困難である。係留ビネット法では自己評定項目に加え、複数の仮想人物について描写した短文 (ビネット) と、それらに対して同様の尺度で評定を求めるビネット項目を導入することにより、両者の分離が可能となる (King et al., 2004)。日本においても係留ビネット法と多次元 IRT モデルを組み合わせたアプローチが提案されており (高橋・岡田, 2016; 北條・岡田, 2017; 北條・岡田, 2018)、健康指標などの調査データに適用されてきた。

これらの研究は、RS を測定したい対象概念とは区別してモデル化し、その分布を推定することで、文化間・集団間の RS の統制を行なっている。

一方、WE の尺度に対して係留ビネット法と多次元 IRT モデルを組み合わせ、RS の影響を統制した研究はこれまで行われていない。

2.5 本研究の位置づけと仮説

既に述べたとおり、日本人の WE スコアが国際的に低いことと回答上のバイアスの関連は繰り返し議論されているものの、バイアスがどの程度スコアを歪めているのかについては十分に検証されていない。本研究ではこの点に対し RS に着目し、係留ビネット法と多次元 IRT モデルを用いた手法を応用することで、RS の WE スコアへの影響を定量的に検討し、新たな知見を示す。このことは、WE 研究だけではなく、他の経営学的な関連指標の測定や実務において応用可能性がある点においても意義があると言える。

以上を踏まえ、本研究では次の仮説を設定する。

H1：日本国内で就業する従業員を対象とした UWES の WE 測定において、回答者の RS はスコアに影響を与える。

H2：RS による WE スコアへの集団レベルの効果は、先行研究の国際比較における日本と他国のスコア差を説明する。

3. 方法

3.1 調査対象者と実施手順

本研究では、アイブリッジ株式会社が運営する調査サービス Freeasy の登録モニターから、日本国内で就業している正規雇用の会社員（18 歳以上 64 歳以下）を対象としてオンライン質問紙調査を実施した（2023 年 11 月 28 日実施）。サンプルが日本の正規職員人口の構成を概ね反映するように、総務省統計（2022）の労働力調査年報に基づき、年齢階級別・性別の人口比に応じて割付を行った。回答は合計 500 名得られたが、全 9 項目に同一選択肢を連続して回答した不誠実回答者を除外し、有効回答 442 名（有効回答率 88.4%）を得た。

3.2 超短縮版 UWES と係留ビネット法の適用

WE の測定には 3 項目で構成される UWES の超短縮版（Ultra-Short Measure; Schaufeli et al., 2019）を用いた。3 項目はそれぞれ 3 つの下位次元と対応し、「全くない」から「いつも感じる／毎日」までの 7 件法により回答を得る。回答者の WE スコアは、3 項目の評定値の平均により算出され、高得点であるほど WE が高いことを示す。UWES にはその他にも 17 項目版や 9 項目版が開発されているが、本研究で 3 項目版を採用した理由の一つは回答者の回答負担の考慮である。一般的に、質問項目が多くなると回答者に疲れや慣れによる疲労効果(fatigue

effect)が生じ、回答が不正確になることが知られている(Choi & Pak, 2005; Krosnick & Presser, 2010)。本研究のように本来の質問項目に加えビネット項目を追加する設計では特にこの問題に注意する必要がある。加えて、局所独立性の前提への配慮も 3 項目版を採用した理由の一つである。IRT 分析では、潜在特性 θ で条件づけたとき項目反応が独立であるという局所独立性の前提がおかれる。Schaufeli et al. (2019)は、9 項目版を用いた確認的因子分析の際に、同一下位次元に属する項目 1 と項目 2 (活力)、および項目 8 と項目 9 (没頭) の各ペアの誤差項に相関を許容しなければモデル適合が得られず、項目ペアが内容的に重複していることを報告している。これは 9 項目版 UWES が局所独立性を十分には満たさない可能性を示唆している。一方で同論文では、3 項目版 UWES は活力・熱意・没頭の各下位次元から 1 項目ずつを選定する設計であり、項目間の誤差相関を許容する必要がなかったことが報告されている。

本研究で用いる係留ビネット法は、2.4 節で既に触れたとおり、自己評定項目とビネット項目の 2 種類のデータが得られる。ビネットから想起される仮想人物の特性の水準が回答者によらず共通であるとすれば、回答者間のビネット項目への回答の違いは回答カテゴリの使い方の違い、すなわち RS のみを反映すると考えることができ、統計モデルを使用して解析を行うことで、回答者の特性と RS を分離した測定が可能となる(北條・岡田, 2018)。係留ビネット法には、測定上の重要な前提として、反応一貫性(response consistency)とビネット等価性(vignette equivalence)がある (King et al., 2004)。前者は、自己評定項目とビネット項目で回答カテゴリの使い方が一貫しているという仮定であり、これが成り立つことによりビネット項目を用いて推定された RS の効果が自己評定にも同じ形で表れていると見なすことができる。後者は、全ての回答者がビネットを同様に解釈し、単一次元上の尺度として評価しているという仮定であり、これが成り立つことにより、ビネット項目の回答に現れる回答者間のばらつきが RS のみを反映していると見なすことができる。本研究では、この仮定と King らの示すビネット項目を設計する上での注意事項をもとに、ビネット項目を設計した。ビネットの人物名には性別に中立的な苗字のみを用いるなど配慮し、活力・熱意・没頭の各次元について 2 つずつ、計 6 つのビネットを作成し、質問紙上では各次元で 2 つのビネット項目を自己評定項目の直後に配置するようにした。仮想人物の描写の面では、自己評定項目の回答分布が 7 件法の中央 (カテゴリ 3) 付近に集中すると想定し、それより低い水準 (1-2 を選びやすい) と高い水準 (4-5 を選びやすい) に位置づく仮想人物を対として設計した。また、収集データに対しビネット等価性の確認を行った。具体的には、回答者の性別 (男性・女性) および年代 (20 代以下・30 代・40 代・50 代以上) の 2 つの属性について、ビネット項目の回答に系統的な差がないかを確認した。性別では Welch の t 検定およびカイ二乗検定を、年代では一元配置分散分析を実施した。有意水準は $\alpha = .05$ とし、多重検定問題の対処として各属性内の検定数 (6 件) に基づき Bonferroni 法を適用した (補正後 $\alpha = .008$)。また、性別の比較には Cohen の d および Cramér の V を、年代の比較には η^2 を算出した。

3.3 データ解析と評価の方法

まず、自己評定項目から算出した WE スコアの記述統計量（平均値・標準偏差）を先行研究で報告されている日本人の WE スコアと比較した。また、3 項目の内的一貫性を確認するため、Cronbach の α 係数を算出した。

次に、北條・岡田（2017）および Bolt et al.（2014）を参考に、自己評定項目とビネット項目の両方を同時に扱う多次元 IRT モデルを設定した。モデルは回答者ごとに測定対象となる潜在特性と RS の 2 次元を想定する構造で、具体的には、自己評定項目について回答者 r が項目 i への回答 U_{ri} としてカテゴリ k を選択する確率を $P(U_{ri} = k | \theta_r, \mathbf{s}_r) = \exp(a_{ik} + s_{rk} + c_{ik}) / \sum_{h=1}^K \exp(a_{ih} + s_{rh} + c_{ih})$ と表現した。ここで、 θ_r は回答者 r の構成概念、つまり今回の場合は測定対象である WE の高さに対応する潜在特性を表すパラメータである。この式のモデルは Bock（1972）によって提案された名義反応モデル(nominal response model)の多次元拡張に位置付けられる(Bolt et al., 2014)。 $\mathbf{s}_r = \{s_{rk}\}$ は回答者 r の各カテゴリ k に対する反応のしやすさを表すパラメータで、回答者 r の RS を反映している。 $\mathbf{a}_i = \{a_{ik}\}$ および $\mathbf{c}_i = \{c_{ik}\}$ はそれぞれカテゴリ間隔と切片を表すパラメータであり、後者は項目 i においてカテゴリ k が選好される度合いを示す。本研究では、自己評定項目について $\mathbf{a}_i = (-3, -2, -1, 0, 1, 2, 3)$ のように隣接カテゴリ間隔を 1 で固定した。これは Bolt et al.（2014）の自己評定項目に対する固定整数値の設定と同様の手法であり、モデルのパラメータ数を抑えて安定的にパラメータ推定を行う役割を持つ。一方、ビネット項目では、評定対象が仮想人物であり回答者自身の潜在特性は反映されないため、 $\mathbf{a}_i = (0, 0, 0, 0, 0, 0, 0)$ と固定し、 θ_r を含めない形にすることで、ビネット項目の情報が反応スタイル $\mathbf{s}_r$ の推定のみに反映される構造となっている。パラメータ推定には、統計解析環境 R（Ver.4.3.1）および RStan（Ver.2.32.3）を用い、マルコフ連鎖モンテカルロ（Markov chain Monte Carlo, 以下 MCMC）法によるベイズ統計解析を実施した。サンプリングの設定は、4 本の独立したチェーン、各チェーン 5,000 回の反復（うち前半 2,500 回をウォームアップとして破棄）とした。収束の判定には、Gelman-Rubin(1992)の収束診断統計量 $\hat{R}$ および有効サンプルサイズを確認した。

また、上述の RS を統制するモデル（統制モデル）とさらに RS を統制しないモデル（非統制モデル）の 2 つを推定し、モデルの比較を行った。非統制モデルでは前述の式で、 $\mathbf{s}_r$ の項を除くことで回答者の RS による影響を考慮しない構造とした。比較には、広く適用可能な情報量規準（Widely-Applicable Information Criterion, 以下 WAIC）および Leave-One-Out 交差検証（以下 LOO）を用いて予測性能を比較した。WAIC および LOO は、その値が小さいほど、予測の観点から優れたモデルであることを示す。

RS を統制したモデルの推定では、各パラメータの事後分布およびそれから得られる事後予測分布を用いることで、効果がどの程度の確率・大きさでどの方向に存在するか直接的に評価を行う。事後予測分布としては、各回答者 r の自己評定項目 i に対する回答カテゴリの予測

値 $\tilde{U}_{ri}$ をサンプリングした。 $\tilde{U}_{ri}$ は推定したモデルから得られる回答データ U_{ri} の予測値である。また、モデル式の s_r を 0 に固定して得られる予測値 $\tilde{U}'_{ri}$ をサンプリングした。 $\tilde{U}'_{ri}$ は回答者の RS の影響を除去した補正後の予測値に相当する。そのうえで、両者の差 $\Delta\tilde{U}_{ri} = \tilde{U}'_{ri} - \tilde{U}_{ri}$ についても事後予測分布を算出した。 $\Delta\tilde{U}_{ri}$ が正の値であれば、回答者 r のスコアは RS によって低く見積もられていたことを意味し、負の値であれば高く見積もられていたことを意味し、その値の大きさは評定尺度上での RS の影響の大きさに対応する。

仮説 H1 の検討においては、各回答者の $\Delta\tilde{U}$ の事後予測分布から、方向性確率 pd (probability of direction; Makowski et al., 2019) と 95% ベイズ信用区間を算出した。 pd は効果の存在の指標であり、 $\Delta\tilde{U} = 0$ の点で事後分布を区切った際の中央値を含む大きい側の面積から求められる。 pd は仮説検定の p 値と対応し、両側検定の場合 $pd = 1 - p/2$ の関係があるとされ、 $pd > .95$ となる回答者については、RS が WE スコアに対して正または負の方向に影響を与えている可能性が高いと判断し、補助的に $\Delta\tilde{U}$ の信用区間が 0 を含むかどうかを確認した。

仮説 H2 については、集団レベルでのバイアス補正効果を評価した。MCMC の各サンプルにおいて、全回答者の事後予測値 $\tilde{U}$ および $\tilde{U}'$ の平均値をそれぞれ算出し、その差 $E[\tilde{U}'] - E[\tilde{U}]$ の事後分布を得た。効果量 (Cohen の d に相当) は、この差の事後平均値を元の WE スコアの標準偏差で除して算出した。効果量の解釈にあたっては、Cohen (1988) の基準を参考にした。また、この補正効果の大きさと先行研究 (Shimazu et al., 2010) で報告されている日本と他国の WE スコアの大まかな差 (約 1 ポイント) を直接比較し、バイアス補正がこの差をどの程度説明しうるかを検討した。

4. 結果

4.1 記述統計

有効回答者 442 名の基本属性を表 1 に示す。平均年齢は 41.7 歳 ($SD = 11.6$, 範囲 18~64 歳) で、性別は男性 291 名 (65.8%), 女性 151 名 (34.2%) であった。婚姻状況は未婚 244 名 (55.2%), 既婚 198 名 (44.8%) で、子供の有無については「なし」が最頻で 276 名 (62.4%) となった。業種は製造業が最も多く (26.7%), その他には「サービス業」や「卸売・小売業」など多様な業種から回答が得られた。居住形態では「持ち家一戸建て」が最も多く 205 名 (46.4%), 都道府県では東京都が最も多く 89 名 (20.1%) となった。

WE スコア測定に関する自己評定項目およびビネット項目の記述統計についても表 1 中に示す。自己評定項目 3 項目から算出した WE スコアの平均値は 2.60 ($SD = 1.41$) であり、先行研究 (Shimazu et al., 2010) で報告されている日本人労働者の WE スコア (2 点台後半) と同程度の水準であった。各項目の平均値は、Q.1 の活力が 2.35 ($SD = 1.63$), Q.4 の熱意が 2.85 ($SD = 1.53$), Q.7 の没頭が 2.60 ($SD = 1.60$) であった。自己評定項目 3 項目の Cronbach の α 係数は、 $\alpha = .86$ であり良好な内的一貫性が確認された。また、WE スコアおよび各変数間の相

関係数を表 2 に示す。WE スコアと各自己評定項目 (Q.1, Q.4, Q.7) との相関は $r = .88 \sim .89$ ($p < .001$) と高く, 3 項目が WE スコアを構成していることと整合的であった。自己評定項目間の相関は $r = .65 \sim .70$ ($p < .001$) であり, 中程度から高い正の相関が認められた。基本属性との関連では, 婚姻状況と WE スコアの間に弱い正の相関 ($r = .12, p < .05$) が認められたが, その他の年齢, 性別, 業種などとの間には有意な相関は認められなかった。

表 1 回答者の基本属性および WE の要約統計量

基本属性 (量的変数)	平均値 (SD)	
年齢	41.7 (11.6)	
基本属性 (質的変数)	カテゴリ (ダミー変数ラベル)	n (割合)
性別	男性 (0)	291 (65.8%)
	女性 (1)	151 (34.2%)
婚姻状況	未婚	244 (55.2%)
	既婚 (1)	198 (44.8%)
子供	なし (0)	276 (62.4%)
	あり (1)	166 (37.6%)
業種	製造業 (0)	118 (26.7%)
	その他 (1)	324 (73.3%)
世帯年収	400~500 万円 (0)	64 (14.5%)
	その他 (1)	378 (85.5%)
居住形態	持ち家一戸建て (0)	205 (46.4%)
	その他 (1)	237 (53.6%)
都道府県	東京都 (0)	89 (20.1%)
	その他 (1)	353 (79.9%)
自己評定項目	平均値 (SD) または値	Cronbach の α
Q. 1 (活力)	2.35 (1.63)	—
Q. 4 (熱意)	2.85 (1.53)	—
Q. 7 (没頭)	2.60 (1.60)	—
WE スコア	2.60 (1.41)	—
内的一貫性	—	.86
ビネット項目	平均値 (SD) または値	
Q. 2 (Q. 1 のビネット)	2.04 (1.46)	
Q. 3 (Q. 1 のビネット)	3.49 (1.50)	
Q. 5 (Q. 4 のビネット)	3.86 (1.56)	
Q. 6 (Q. 4 のビネット)	1.91 (1.35)	
Q. 8 (Q. 7 のビネット)	3.73 (1.52)	
Q. 9 (Q. 7 のビネット)	2.71 (1.24)	

注：質的変数についてはダミー変数として最頻値を 0, その他のカテゴリをまとめて 1 とした。

注：WE スコアは 3 つの自己評定項目 (Q. 1, Q. 4, Q. 7) の平均値。

表 2 WE スコアおよび質問項目、各属性の相関係数行列

	WE	Q.1 活力	Q.4 熱意	Q.7 没頭	年齢	都道府県	性別	結婚	業種	年収	居住
Q.1 活力	.88***										
Q.4 熱意	.89***	.69***									
Q.7 没頭	.88***	.65***	.70***								
年齢	-.01	-.02	.03	-.03							
都道府県	-.02	-.07	-.01	.03	.01						
性別	-.03	-.01	-.01	-.07	-.11*	-.07					
結婚	.12*	.11*	.10*	.10*	.28***	.04	-.12*				
業種	.06	.05	.03	.07	-.02	-.05	.17***	0			
年収	.03	.02	0	.06	.01	-.01	.07	.14**	.07		
居住	-.08	-.02	-.11**	-.09	-.09*	-.18***	.07	-.09	.12*	-.03	
子供	.09	.09	.10*	.06	.27***	.10*	-.12*	.72***	0	.12*	-.20***

注： * $p < .05$, ** $p < .01$, *** $p < .001$

ビネット等価性の確認については、3.2 節で述べた方法により検定を実施した結果、有意水準を下回る p 値を示したのは年代に関する 2 件であった。性別による比較ではいずれの検定も 6 項目すべてにおいて有意水準を下回らず、効果量も最大で $d = 0.21$, $V = 0.19$ と小さい水準にとどまった。年代では 6 項目のうち、Q.6（熱意：低 WE ビネット）において $F(3, 438) = 8.71, p < .001, \eta^2 = .056$, Q.9（没頭：低 WE ビネット）において $F(3, 438) = 4.39, p = .005, \eta^2 = .029$ と、2 つの項目で有意な差が認められた。事後的に Welch の t 検定（Bonferroni 補正, $\alpha = .008$ ）を行ったところ、Q.6 では 20 代以下が 40 代（ $p < .001$ ）および 50 代以上（ $p < .001$ ）よりも有意に高い評定を示し、Q.9 では 40 代が他の 3 群と比較して有意に低い評定を示した。Q.6 と Q.9 はいずれも WE が低い仮想人物を描写した項目であり、管理職や主任といった職位にある人物が業務上の困難を抱えている状況が描かれている。同様に職位に関する記述を含む Q.3（高 WE ビネット）では年代差が認められなかったことから、職位の記述そのものではなく、管理職としての困難を描写した内容について、その職位の経験が多い 40 代以上の回答者がより深刻に評価した可能性がある。この解釈差はビネット等価性の部分的な違反に該当する可能性があるが、一方で年代が説明する分散の割合は高々 $\eta^2 = .056$ であり、Cohen (1988) の基準（小 = .01, 中 = .06, 大 = .14）に照らすと中程度の閾値（.06）を下回っている。したがって、この結果がビネット等価性の前提を大きく損なうものではないと判断した。

4.2 モデル推定結果の比較

MCMC サンプリングを行ったのち、 $\hat{R}$ が全てのモデル・パラメータについて 1.1 未満となることを確認し収束したと判断した。また、有効サンプルサイズもすべてのパラメータで 100 以上であることを確認した。回答者の RS を統制したモデルと非統制モデルの比較を行うため、算出した WAIC および LOO を表 3 に示す。活力・熱意・没頭のいずれの質問項目において

も、統制モデルがより低い値を示し、RS を考慮したモデルの方がデータへの当てはまりが良いことが確認された。以降の解析では RS を統制したモデルを使用した。

表 3 RS を統制したモデルと非統制モデルの情報基準量の比較

項目 (概念)	WAIC (SE)		LOO (SE)	
	統制	非統制	統制	非統制
Q. 1 (活力)	113.06 (16.15)	404.15 (9.85)	114.15 (15.95)	403.94 (9.83)
Q. 4 (熱意)	160.65 (14.16)	376.65 (9.90)	160.72 (14.06)	376.57 (9.90)
Q. 7 (没頭)	150.94 (14.24)	383.99 (10.22)	150.89 (14.17)	383.83 (10.21)

4.3 RS 統制モデルの推定結果の詳細

RS を統制したモデルにおける、各回答者の $\Delta\hat{U}$ の点推定値（事後平均）については、その絶対値のサンプル全体での平均を求めたところ、Q.1 の活力で 0.25 ($SD=0.48$, 中央値 = 0.06), Q.4 の熱意で 0.16 ($SD=0.32$, 中央値 = 0.05), Q.7 の没頭で 0.18 ($SD=0.36$, 中央値 = 0.05) となった。点推定値の大きさが 0.5 以上とカテゴリを跨ぐ可能性が比較的大きい回答者の割合をみると、Q.1 の活力で 12.9%, Q.4 の熱意で 7.5%, 没頭 9.0%であった。また、統制モデルにおける自己評定項目の切片 c_i の事後分布について確認を行った際の事後平均値および 95% 信用区間を表 4 に示す。

表 4 RS 統制モデルにおける自己評定項目の切片 $c_{i,k}$ の事後平均と 95% 信用区間

項目 (概念)	k=0	k=1	k=2	k=3	k=4	k=5	k=6
Q. 1 (活力)	-0.36	+0.96	+1.54	+1.71	+0.24	-0.91	-3.19
	[-0.84, +0.11]	[+0.60, +1.32]	[+1.24, +1.83]	[+1.46, +1.98]	[-0.11, +0.59]	[-1.35, -0.48]	[-3.92, -2.49]
	-1.66	+0.09	+1.26	+1.87	+1.02	-0.34	-2.24
Q. 4 (熱意)	[-2.14, -1.17]	[-0.28, +0.44]	[+0.97, +1.55]	[+1.63, +2.11]	[+0.72, +1.32]	[-0.74, +0.03]	[-2.84, -1.71]
	-0.76	+0.35	+1.53	+1.61	+0.53	-0.61	-2.66
	[-1.24, -0.28]	[-0.02, +0.74]	[+1.26, +1.82]	[+1.35, +1.86]	[+0.21, +0.85]	[-1.02, -0.22]	[-3.31, -2.05]

$\Delta\hat{U}$ の事後予測分布から求めた、方向性確率 $pd \leq .95$ となる回答者および、95%信用区間が 0 を含む回答者、事後平均値が正となる回答者が全体に占める割合を表 5 に示す。項目全体において、 $pd \leq .95$ となる回答者の割合は 99.8%であり、過半数を大きく上回った。また、95%信用区間が 0 を含む回答者の割合は 100.0%であった。活力・熱意・没頭の各下位概念においても、 $pd \leq .95$ の割合および 95%信用区間が 0 を含む割合はいずれも 100.0%であった。

表 5 $\Delta \tilde{U}$ の事後分布をもとに個人レベルのバイアスを集計した結果

概念	$pd \leq .95$ の回答者の割合	95%信用区間が 0 を含む回答者の割合	事後平均値>0 の回答者の割合
活力	100.0%	100.0%	45.2%
熱意	100.0%	100.0%	48.4%
没頭	100.0%	100.0%	46.4%
WE スコア	99.8%	100.0%	45.9%

また、バイアス補正の集団レベルの効果を確認するため、Cohen の d を算出した結果を表 6 に示す。項目全体において、 $E[\tilde{U}'] - E[\tilde{U}]$ の事後平均は 0.013 であり、95%信用区間(-0.070, 0.097)が 0 を含んでいた。Cohen の d は 0.010 であり、Cohen (1988) の基準に照らすと極めて小さい効果であった。また、この補正効果は国際比較差(約 1 ポイント)の 1.3%に相当した。下位概念別にみても、活力 ($d=0.029$)、熱意 ($d=0.003$)、没頭 ($d=-0.007$) のいずれも実質的な意味を持たない水準であった。これらの結果の解釈については次章で考察する。

表 6 集団レベルでのバイアスの効果量

概念	$E[\tilde{U}'] - E[\tilde{U}]$ の事後平均	95%信用区間	Cohen の d	国際比較差に対する 事後平均の割合
活力	0.047	[-0.100, 0.195]	0.029	4.7%
熱意	0.005	[-0.136, 0.143]	0.003	0.5%
没頭	-0.011	[-0.158, 0.138]	-0.007	1.1%
WE スコア	0.013	[-0.070, 0.097]	0.010	1.3 %

注：国際比較差に対する割合は、Shimazu et al. (2010) で報告された日本と 15 カ国平均の差をもとに 1 ポイントを基準として算出した。

さらに、WE 水準の高低によって RS に違いが現れるかどうかの確認も行った。具体的には、活力・熱意・没頭の各下位概念において、回答者の潜在特性 θ の点推定値(事後平均)が $\theta > 0$ を高群、 $\theta \leq 0$ を低群として群分けを行い ($\theta = 0$ の回答者は存在せず、高群/低群のサイズは活力 202/240, 熱意 262/180, 没頭 227/215 となった)、各回答者の RS の大きさとして $\Delta \tilde{U}_i$ を比較した。なお、各概念とも θ の事後平均と回答スコアの相関は $r > .94$ と極めて高く、 θ による群分けは観測スコアによる分割と実質的に同義とみなせると判断した。 $\Delta \tilde{U}_i$ の事後平均値について高群/低群で群平均を求めたところ、活力で -0.201/+0.256, 熱意で -0.104/+0.163, 没頭で -0.163/+0.149 と符号の反転が確認された。一方 $|\Delta \tilde{U}_i|$ の事後平均値の群平均は、活力で 0.215/0.277, 熱意で 0.136/0.193, 没頭で 0.182/0.173 と差は小さく、Cohen の d の絶対値は活力 0.128, 熱意 0.175, 没頭 0.027 で、いずれも Cohen (1988) の基準における「小さい効果」($d=0.2$) に満たない水準であった。

5.考察

5.1 本研究の結果の要約

仮説 H1「日本国内で就業する従業員を対象とした UWES の WE 測定において、回答者の RS はスコアに影響を与える」については、WAIC と LOO を用いたモデルの比較で RS を統制したモデルが統制しなかったモデルに比べ予測の観点から良いモデルであると示されたことから、RS の存在という観点ではスコアへの影響が部分的に支持された。しかし一方で、 $\Delta\tilde{U}$ の事後予測分布から求めた方向性確率 pd が .95 を超える回答者の割合がわずか 0.2%にとどまり、95% 信用区間が 0 を含む回答者の割合が 100.0%であったことや、 $\Delta\tilde{U}$ の点推定値の大きさが半カテゴリを超えるような回答者の割合は最大でも約 13%と限定的な水準であったことなどを踏まえると、回答者の RS が WE スコアに対しそのスコアが変化する程の実質的な影響を与えているとは言えず、今後の検証の余地が残る。

仮説 H2「回答バイアスによる WE スコアへの集団レベルの効果は、先行研究の国際比較における日本と他国のスコア差を説明する」については、補正効果 ($E[\tilde{U}'] - E[\tilde{U}]$) の事後平均が 0.013 ポイント (95% 信用区間: -0.070, 0.097) , Cohen の d が 0.010 と極めて小さい値であったことから、支持されなかった。この補正効果は、先行研究で報告されているおおよその国際比較差 1 ポイントを基準とした時の 1.3%にすぎない値であった。

また、WE の水準別の比較では、潜在特性 θ の高群と低群で $\Delta\tilde{U}_i$ の点推定値の平均を求めると、符号の反転が確認された。このことは、それぞれの群において異なる RS の傾向を示唆するものの、 $|\Delta\tilde{U}_i|$ の群平均には実質的な差が見出されず、また個人レベルで信用区間が 0 を含まない回答者は認められなかったことから、この傾向が WE スコア平均値を押し下げる要因となるとは考えにくい。

他方、項目切片 $\mathbf{c}_i$ の事後分布では、自己評定の 3 項目全てにおいて最高カテゴリ ($k = 6$) の事後平均が大きく負、中央付近のカテゴリで強く正となるパターンが観察された。最低カテゴリ ($k = 0$) でも値が負となる傾向はみられるが、3 項目とも最高カテゴリを避ける傾向が明らかに強く、Shimazu et al. (2010) に言及される日本人のポジティブな感情の表出を抑制する傾向 (Iwata et al., 1995) と整合する傾向である。ただし、 $\mathbf{c}_i$ は項目内容に依存するパラメータであり、項目内容に依存しないとされる RS とは区別される。項目内容に依存するようなバイアスとしては社会的望ましき反応などが挙げられ (Rorer, 1965; 藤岡, 2023) , その反応が $\mathbf{c}_i$ の推定結果に反映された可能性が考えられる。

5.2 学術的意義

本研究の学術的意義は三点ある。第一に、日本の WE スコア水準をめぐるバイアスによる影響の議論を、RS に着目しその影響を統制することで、直接的かつ定量的な検証を試みた点である。この点で本研究は、推測や間接的な評価にとどまっていたこれまでの議論を、定量的に検証した数少ない有用な知見になると考えられる。第二に、RS による WE スコアへの集団

レベルでの効果は、先行研究の国際比較で報告される日本と他国とのスコア差を実質的に説明しないことを示し、「日本人は控えめに回答する RS を持つためWEスコアが低く見える」という説明では、日本と他国との平均スコアの差を説明できないことを示した点である。一方で、5.1 節で言及したように、個々の回答者に依存する RS とは異なる、社会的望ましき反応といった、項目内容に依存する反応が WE スコアに影響を与えている可能性も示唆された。これらの知見は、今後の研究に対して次の二つの重要な方向性を示唆する。一つは、国際比較を通じて、UWES の項目内容が回答者集団のカテゴリ選好に与える影響に着目し、その国家間の違いを検証することである。もう一つは、こうした測定上の検証と並行して、WE スコアの水準差を生み出す本質的な要因を探求することである。後者については、Job Demands-Resources モデルの先行要因である仕事の資源（裁量権、フィードバック、上司・同僚からの支援、成長機会など）や、個人の資源（自己効力感・レジリエンス）などを、他国との違いと併せて捉え直し、WE スコアが低水準である本質的な要因を検討していくことが重要である。特に日本の職場においては管理職のマネジメント・スキルが不足している可能性も指摘されており（岩澤, 2016）、職務裁量やフィードバック、成長機会の提供といった仕事の資源の設計、さらには組織のマネジメント構造との関連も視野に入れ今後議論していくことが求められる。第三に、本研究で使用した係留ビネット法と多次元 IRT モデルを組み合わせた分析の枠組みは、WE だけでなく他の経営学的な関連指標の測定にも応用が可能であるという点で、本研究がその枠組みの導入的事例となることに意義がある。例えば、WE 同様ポジティブな心理状態に関連する指標の「仕事の満足度」や「組織コミットメント」についても米国と比べて日本人の水準が低いことが報告されており（Lincoln & Kalleberg, 1985）、日本人特有の回答バイアスが十分に補正されていないという可能性が指摘されている（岩澤, 2016）。本研究で用いたアプローチは、WE 尺度のみならず、これらの指標にも十分適用可能であり、経営学的指標の国際比較研究に内在するバイアスの問題に対して汎用的な解決策を提示するものといえる。

5.3 実践的意義

本研究の結果は、日本における WE スコアの低さが測定上のバイアスに起因するという従来の議論に対し、RS という観点からはその影響が限定的である可能性を示すものである。したがって、スコアの低さを RS に起因するものとして捉えるのではなく、日本の企業や組織においては、「なぜ従業員のエンゲイジメントが低いのか」という問いを正面から受け止め、その原因を実務的な視点から丁寧に探索することが求められる。たとえば、職場環境や業務設計、上司との関係性、キャリアの見通しといった具体的な要因を組織内で検討・改善していくことが、WE 向上に向けた実効性ある施策の出発点となりうる。

6. おわりに

本研究には、いくつかの限界がある。第一に、本研究で用いたデータは、インターネット調査会社のモニターを対象としており、日本の正規雇用者全体を完全に代表するものではない

点である。特定の属性の回答者の偏りや、インターネット調査特有の選択バイアスが存在する可能性を否定できないため、結果の再現性の検証が今後の課題となる。第二に、本研究の主要な関心は回答者の RS に関連するパラメータを通じて WE スコアへの影響を推定することにあるが、潜在特性 θ_r 自体の推定精度という観点では、3 項目という最低限の項目数でさらに下位次元ごとに 1 項目しか用いないことに起因する制約が存在する。多値 IRT では一般に項目数が多くなるほど情報量が加法的に増加し θ_r の推定精度が高まるため、より多くの項目数からなる UWES を用いた調査や、WE 全体を単一の潜在特性として扱う、あるいは 3 つの下位次元を多次元的に扱うことで全ての自己評定項目から同時にパラメータを推定するモデルの検討が今後の課題となる。ただしその場合は、局所独立性や下位次元の構造に関する既存の知見 (De Bruin et al., 2013; Schaufeli et al., 2019) を踏まえた慎重な検討が必要である。第三に、本研究のモデルは、Bolt et al. (2014) および北條・岡田 (2017, 2018) において採用されている設計に倣い、自己評定項目の α_i をカテゴリ間隔が 1 となるように固定したが、この操作は安定したパラメータ推定が可能となる一方で、識別力に関してモデルの柔軟性を損ない、パラメータの推定に影響を及ぼしている可能性も考えられる。今後この点を検証するために、識別力すなわちカテゴリ間隔の値についても推定を行うモデルを検討し、モデルの比較を行うことが課題となる。第四に、本研究は UWES により取得した国内の WE データのみを用いており、RS と WE の関係についてより頑健な結論を得るためには、他国データとの直接的な比較や、仕事満足度や組織コミットメントといった他の関連概念の尺度との比較が今後の課題となる。特に、WE スコアの国際比較の研究について Iwata et al. (1995)は、「欧米ではポジティブな感情を積極的に表出することが望ましいとされている」ことを指摘していることから、欧米の国々においても RS を統制した同様のアプローチの適用を行ったうえで日本のデータと比較することが望まれる。第五に、本研究では係留ビネット法を WE スコアの測定に応用した導入的な位置付けであるため、自己評定 1 項目に対してビネット 2 項目という最小限の構成を採用した。RS パラメータの推定精度を高めるためには、ビネット項目数を増やし RS パラメータの情報量を増加させることが今後の調査では重要である。第六に、係留ビネット法的前提であるビネット等価性については、4.1 節で性別・年代による比較を行ったが、年代に関して 2 項目で部分的な違反の可能性が認められた。これらの項目では、課題を抱えた管理職の描写が、回答者の職位経験を通じ年代による解釈差として現れた可能性がある。今後新たに同様の手法を用いる際は、特定の職位や役割に依存しない場面設定を用いるなどの工夫により、回答者の属性による解釈差が生じにくいビネットの設計が重要となる。一方で、反応一貫性の仮定については本研究の中で検証を行っておらず、その実証的検証手続きの検討や適用は今後の重要な課題である。第七は、本研究の結果を直ちに日本人および日本の文化に固有の特性として一般化することはできない点である。本研究は先行研究(Shimazu et al., 2010)において日本人特有の文化的バイアスとして指摘されてきた回答傾向を、RS の一形態として捉え、RS の影響を検証する設計を採用した。そのため、本研究では日本と他国の文化的

背景の違いを直接比較する設計を採用しておらず、例えば Hofstede (1991)の文化次元に代表されるような文化指標の測定は行っていない。また、調査対象は日本国内で就業する正規雇用者に限定しており、回答者の国籍の確認は行っていない。したがって、RS の影響や WE の水準が日本人またはその文化に特有のものかどうかを判断するためには、国籍の確認や文化に関する指標の測定を行ったうえで、日本と他国のサンプルを直接的に比較する研究が今後必要である。

【文献リスト】

岩澤誠一郎 (2016). 「日本企業における従業員のワーク・エンゲイジメントとマネジメント・スキル」 経済社会学会年報, 38, 72-90.

窪田和巳・島津明人・川上憲人 (2014). 「日本人労働者におけるワーカホリズムおよびワーク・エンゲイジメントとリカバリー経験との関連」 行動医学研究. 20(2), 69-76.

島津明人 (2010). 「職業性ストレスとワーク・エンゲイジメント」 ストレス科学研究. 25, 1-6.

島津明人 (2014). 『ワーク・エンゲイジメント: ポジティブ・メンタルヘルスで活力ある毎日を』, 労働調査会.

島津明人 (2025). 「ワーク・エンゲイジメント: 研究と実践における最近 10 年の動向」 産業医学レビュー. 37(3), 121-142.

高橋雄介・岡田謙介 (2016). 「調査データの回答バイアスの補正方法としての係留寸描法の有効性」 人工知能学会全国大会論文集. 30, 3B3NFC05a2-3B3NFC05a2.

田崎勝也・申知元 (2017). 「日本人の回答バイアス——レスポンス・スタイルの種別間・文化間比較——」 心理学研究. 88(1), 32-42.

藤岡慧 (2023). 「社会的望ましさを考慮した項目反応モデルによるパラメタ推定法の提案」 心理学研究. 94(4), 300-307.

北條大樹・岡田謙介 (2017). 「評定尺度における反応傾向を考慮した係留寸描データのベイズ的項目反応モデル」 データ分析の理論と応用. 6(1), 113-125.

北條大樹・岡田謙介 (2018). 「係留ビネット法による反応スタイルの分類 ヨーロッパの大規模健康調査を例に」 行動計量学. 45(1), 13-25.

米村恵子・松田義幸 (1991). 「社会調査に観る日本人の曖昧・中間回答に関わる研究」 日本社会心理学会第 32 回大会発表論文集, 66-69.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1, Pt. 1), 29-51.

- Bolt, D. M., Lu, Y., & Kim, J. S. (2014). Measurement and control of response styles using anchoring vignettes: a model-based approach. *Psychological methods*, 19(4), 528.
- Chen, C., Lee, S. Y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological science*, 6(3), 170-175.
- Choi, B. C., & Pak, A. W. (2005). A catalog of biases in questionnaires. *Preventing chronic disease*, 2(1), A13.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.
- De Bruin, G.P., Hill, C., Henn, C.M., & Muller, K-P. (2013). Dimensionality of the UWES-17: An item response modelling analysis. *SA Journal of Industrial Psychology/SA Tydskrif vir Bedryfsielkunde*, 39(2), 1–8.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4), 457-472.
- Hofstede, G. (1991). *Cultures and Organizations: Software of the Mind*. McGraw-Hill.
- Iwata, N., Roberts, C. R., & Kawakami, N. (1995). Japan-US comparison of responses to depression scale items among adult workers. *Psychiatry Research*, 58(3), 237-245.
- Johnson, T. P., Shavitt, S., & Holbrook, A. L. (2011) . Survey response styles across cultures. In D. Matsumoto & F. J. R. van de Vijver (Eds), *Cross-cultural research methods in psychology*. New York: Cambridge University Press. pp. 130- 175.
- Kahn, W. A. (1990). Psychological conditions of personal engagement and disengagement at work. *Academy of management journal*, 33(4), 692-724.
- Kankaraš, M., & Moors, G. (2011). Measurement equivalence and extreme response bias in the comparison of attitudes across Europe: A multigroup latent-class factor approach. *Methodology*, 7, 68–80.
- King, G., Murray, C. J., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American political science review*, 98(1), 191-207.
- Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research*. 2nd (pp. 263–313). Bingley, UK: Emerald Publishing Group.
- Lincoln, J. R., & Kalleberg, A. L. (1985). Work organization and workforce commitment: A study of plants and employees in the US and Japan. *American Sociological Review*, 50(6), 738-760.
- Makowski, D., Ben-Shachar, M. S., Chen, S. A., & Lüdtke, D. (2019). Indices of effect existence and

significance in the Bayesian framework. *Frontiers in psychology*, 10, 2767.

Mazzetti, G., Robledo, E., Vignoli, M., Topa, G., Guglielmi, D., & Schaufeli, W. B. (2023). Work Engagement: A meta-Analysis Using the Job Demands-Resources Model. *Psychological reports*, 126(3), 1069–1107.

Rorer, L. G. (1965). The great response-style myth. *Psychological Bulletin*, 63(3), 129-156.

Schaufeli, W. B., Salanova, M., González-Romá, V., & Bakker, A. B. (2002). The measurement of engagement and burnout: A two sample confirmatory factor analytic approach. *Journal of Happiness studies*, 3(1), 71-92.

Schaufeli, W. B., & Bakker, A. B. (2004). Job demands, job resources, and their relationship with burnout and engagement: A multi-sample study. *Journal of Organizational Behavior*, 25(3), 293-315.

Schaufeli, W. B., Shimazu, A., Hakanen, J., Salanova, M., & De Witte, H. (2019). An ultra-short measure for work engagement: The UWES-3 validation across five countries. *European Journal of Psychological Assessment*, 35(4), 577–591.

Shimazu, A., Schaufeli, W. B., Kosugi, S., Suzuki, A., Nashiwa, H., Kato, A., ... & Kitaoka-Higashiguchi, K. (2008). Work engagement in Japan: validation of the Japanese version of the Utrecht Work Engagement Scale. *Applied Psychology*, 57(3), 510-523.

Shimazu, A., Schaufeli, W. B., Miyanaka, D., & Iwata, N. (2010). Why Japanese workers show low work engagement: An item response theory analysis of the Utrecht Work Engagement scale. *BioPsychoSocial Medicine*, 4(1), 17.

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International journal of public opinion research*, 25(2), 195-217.

Yang, Y., Harkness, J. A., Chin, T. Y., & Villar, A. (2010). Response styles and culture. *Survey methods in multinational, multiregional, and multicultural contexts*, 203-223.

Mao SAWADA[†]

Hiroya HIRAKIMOTO[‡]

Abstract

This study aimed to quantitatively examine the extent to which response style (RS) — a bias arising at the time of responding and previously discussed as a possible source of cultural bias — contributes to the internationally low work engagement (WE) scores of Japanese workers. A questionnaire survey was conducted with 442 regular employees recruited through an internet survey company.

The survey included three items from the Ultra-Short Measure of the Utrecht Work Engagement Scale (UWES) and six anchoring vignette items. Bayesian analysis using a multidimensional item response theory model combined with the anchoring vignette method was employed to estimate a model controlling for respondents' RS. The results showed that, although model comparison suggested the presence of RS effects, no RS of meaningful magnitude was identified for nearly all respondents, and RS did not substantively affect their WE scores. Furthermore, the correction effect from controlling for RS was extremely small at 0.013 points, representing only 1.3% of the approximately one-point difference reported in previous international comparisons.

The effect size (Cohen's $d = 0.010$) was also negligible. These findings suggest that the low WE scores in Japan are difficult to explain by RS, and that other potential influences — including biases that depend on item content, as well as job characteristics and the work environment — warrant further examination.

JEL code : O15 J24 J28

Keywords : Work Engagement, Utrecht Work Engagement Scale, Response Style, Anchoring Vignette Method, Multidimensional IRT Model

[†] Graduate Student, Graduate School of Economics, The University of Osaka
mail:u338546k@ecs.osaka-u.ac.jp

[‡] Graduate School of Economics, The University of Osaka mail: hirakimoto@econ.osaka-u.ac.jp