An investigation of the item length in the forced-choice psychological measurement

Kyosuke Bunji
Kobe University

Takeshi Sugiyama
Recruit Management Solutions Co., Ltd.

Kensuke Okada
University of Tokyo

✉ bunji@bear.kobe-u.ac.jp

The 13th Conference of the IASC-ARS
Dec 4, 2025

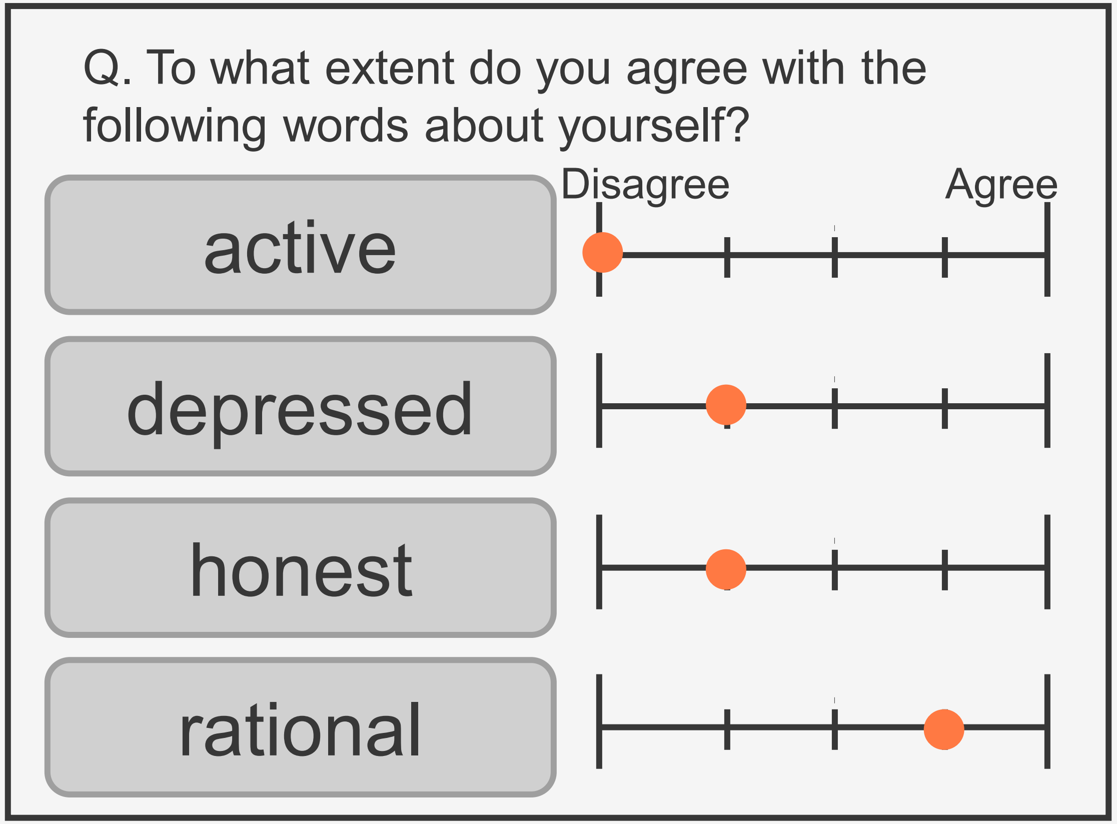
1

Introduction

Forced-Choice (FC) scale

Respondents are required to **choose the best option** in each block

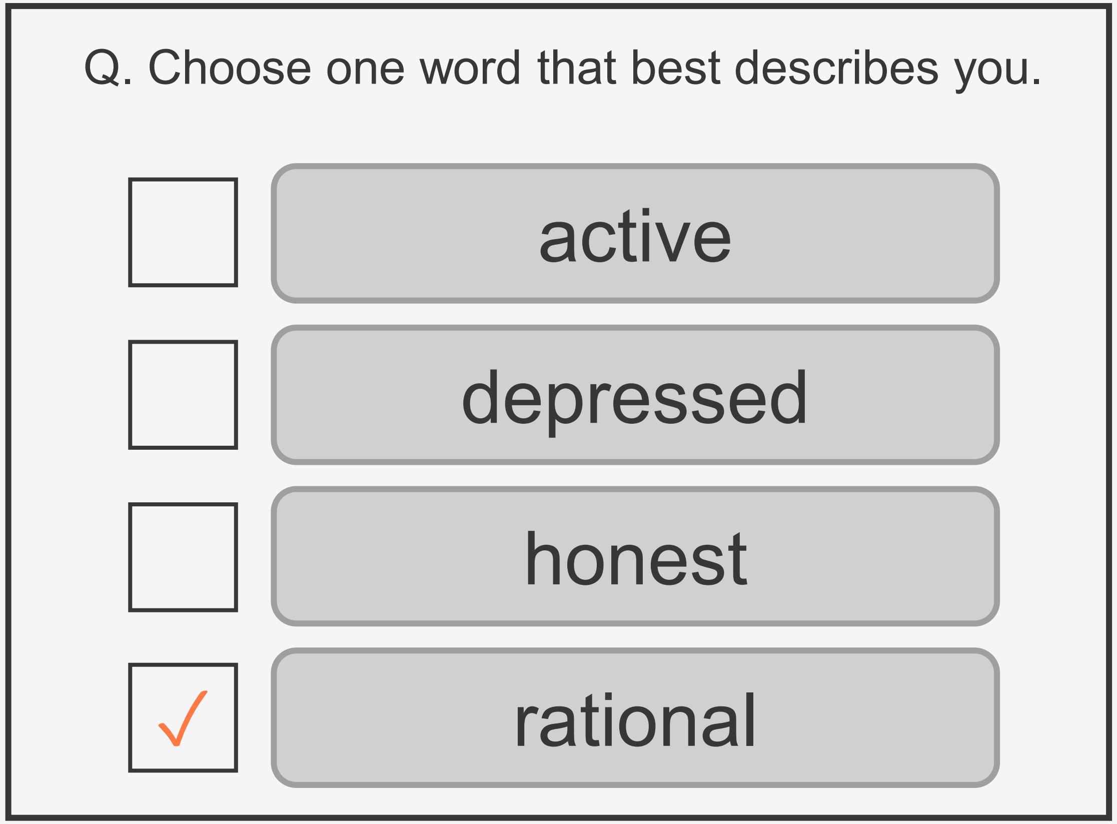
Single-Stimulus (SS); Likert Scale



Q. To what extent do you agree with the following words about yourself?

... is frequently contaminated by systematic response biases

(4-Alternative) Forced-Choice



Q. Choose one word that best describes you.

□ active

□ depressed

□ honest

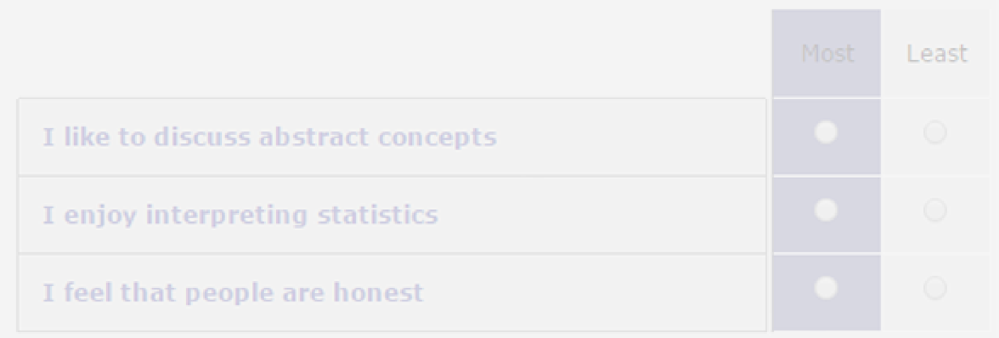
☑ rational

... is designed to reduce systematic response biases

Concern about the FC scale

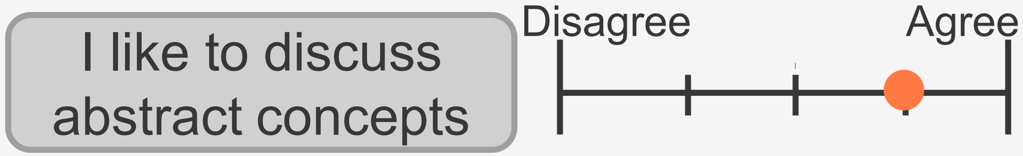
It would be difficult to answer each block

Ex) OPQ32
（One of the most famous assessment tools for job conduct）

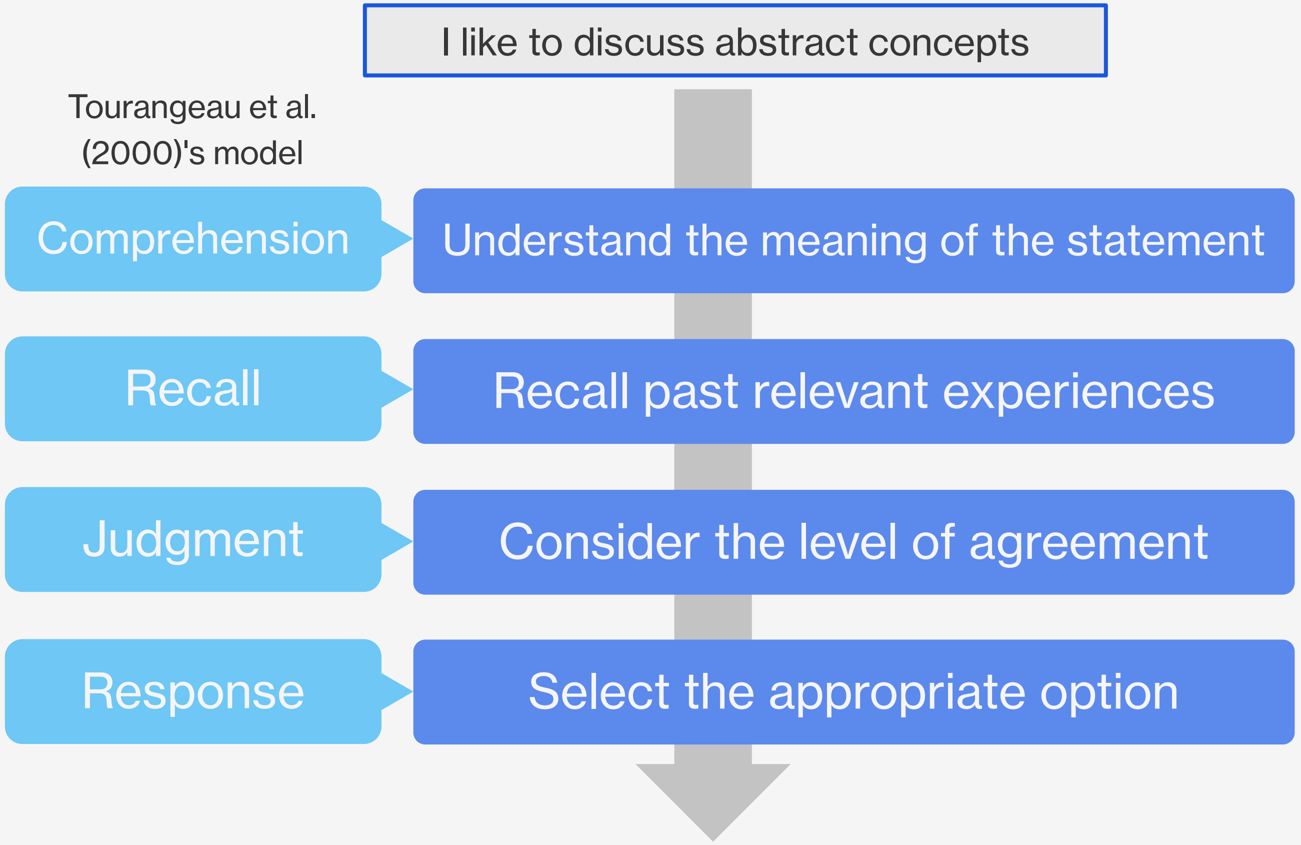


	Most	Least
I like to discuss abstract concepts	○	○
I enjoy interpreting statistics	○	○
I feel that people are honest	○	○

Q. To what extent do you agree with the following words about yourself?



In the case of Likert format



I like to discuss abstract concepts

Tourangeau et al. (2000)'s model

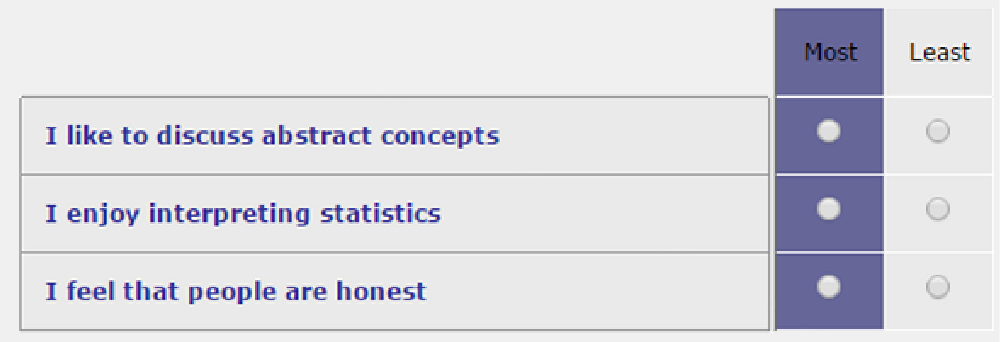
- Comprehension → Understand the meaning of the statement
- Recall → Recall past relevant experiences
- Judgment → Consider the level of agreement
- Response → Select the appropriate option

"Somewhat agree"
I guess that's about it!

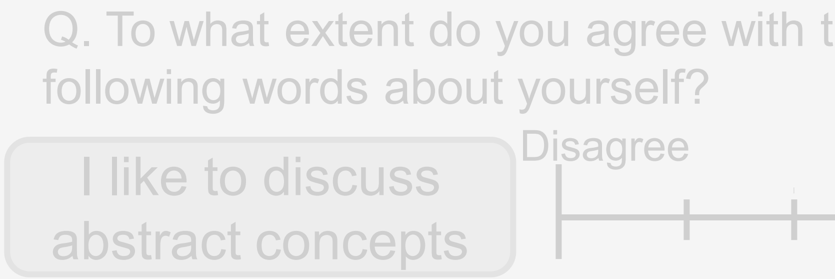
Concern about the FC scale

It would be difficult to answer each block

Ex) OPQ32
（One of the most famous assessment tools for job conduct）

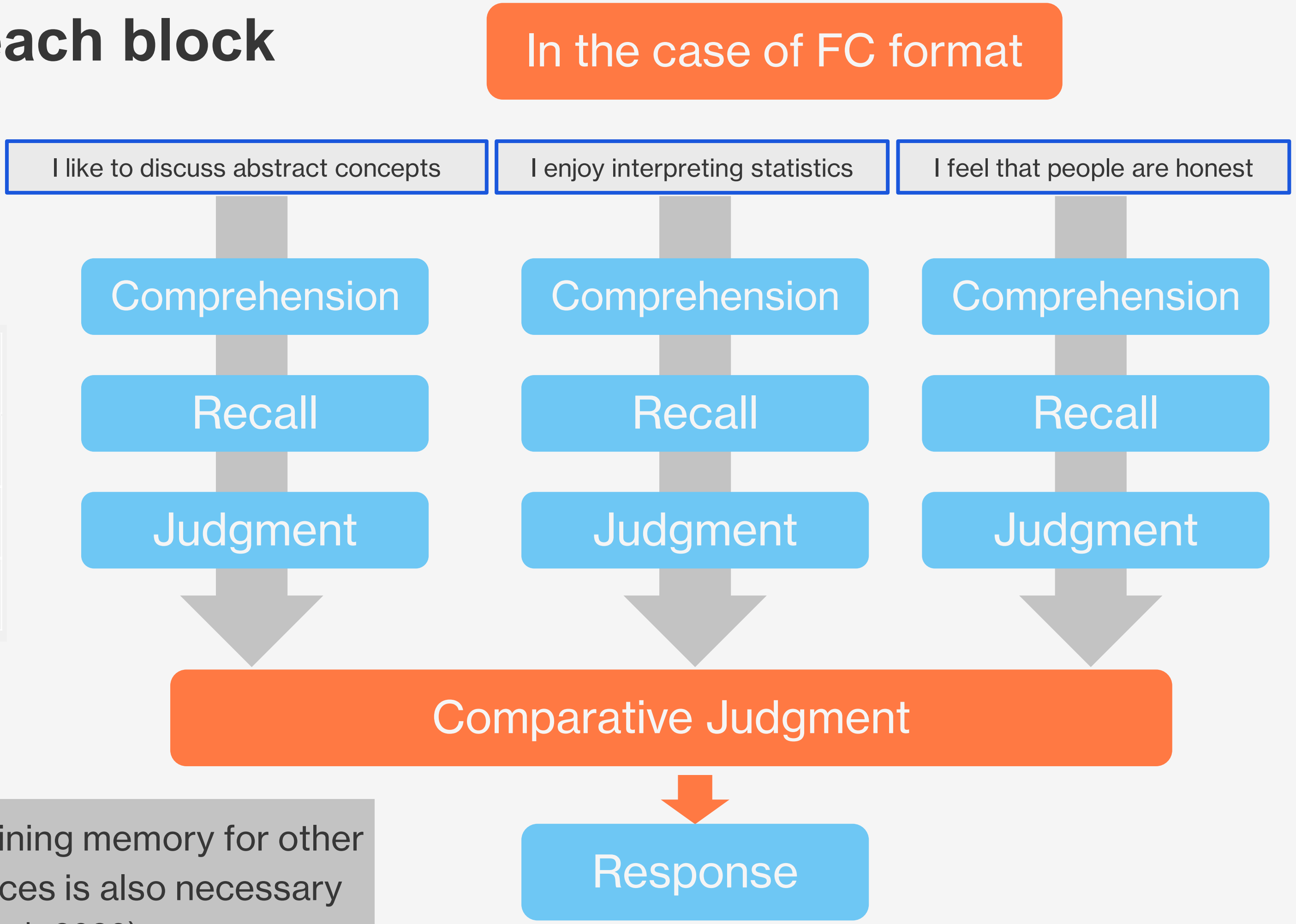


	Most	Least
I like to discuss abstract concepts	●	○
I enjoy interpreting statistics	●	○
I feel that people are honest	●	○



Q. To what extent do you agree with the following words about yourself?

Maintaining memory for other sentences is also necessary (Sass et al., 2020)



I enjoy interpreting statistics

may be the best one... mmm....

Statement (item) length affects the difficulty

Lenzer et al. (2010)

Lengthy or complex questions may overload one's working memory

Ex) a sentence requires readers to hold a lot of information, or includes many logical operators (and/or)

Alwin & Beattie (2016)

Reliability decreases as the number of words in a question increases

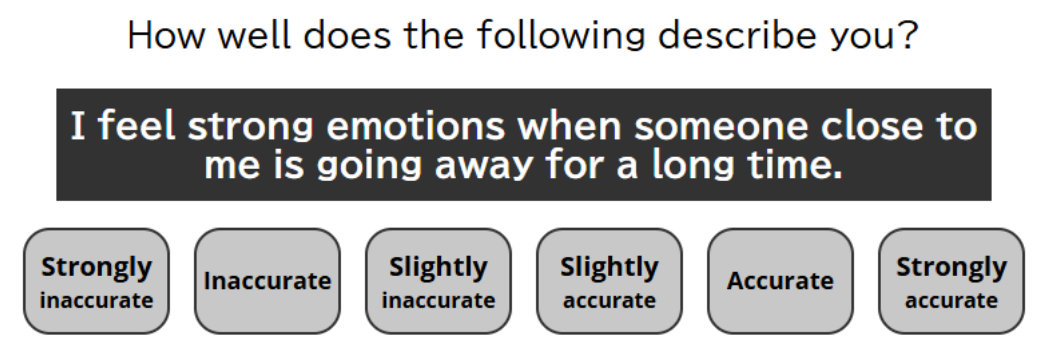
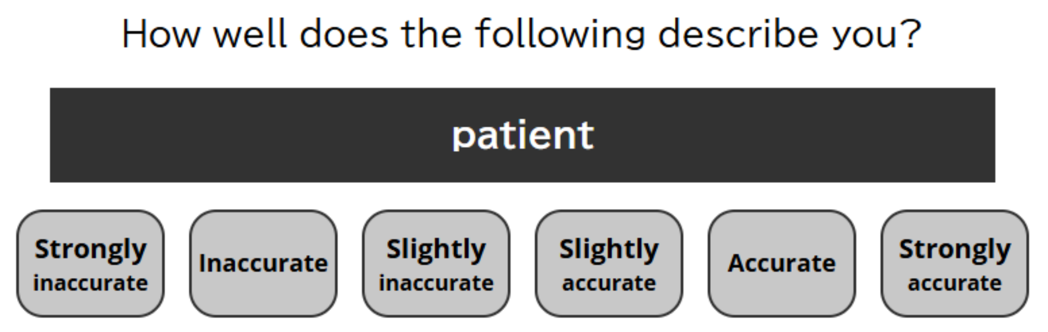
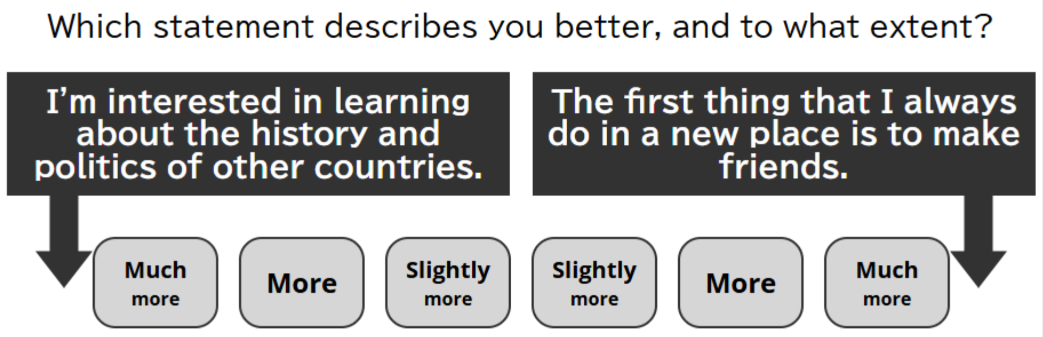
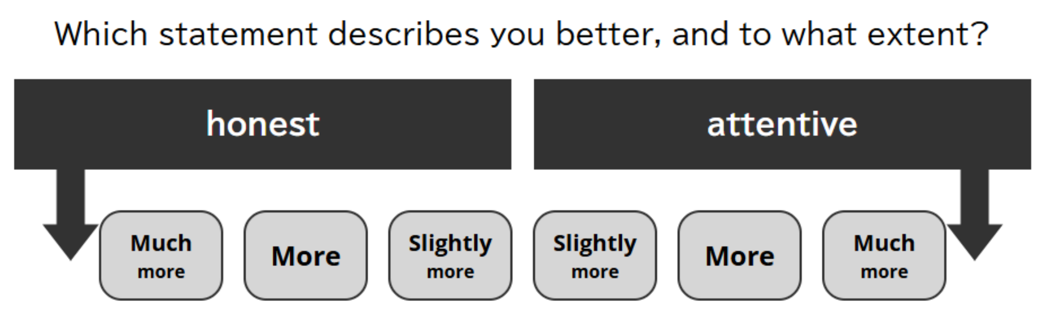
➔ KISS (Keep it simple, stupid) principle should be observed

Hamby and Ickes (2015)

Shorter and more "de-contextualized" items show better performance (Cronbach's α) in terms of assessing personality traits

Longer items may seem to have adverse effects to response quality

Screen image

	HEXACO (Ashton & Lee, 2009)	HEXACO adjective scale (Romano et al., 2023)
	(Short) sentence	Adjective
Likert	 How well does the following describe you? I feel strong emotions when someone close to me is going away for a long time. Strongly inaccurate Inaccurate Slightly inaccurate Slightly accurate Accurate Strongly accurate	 How well does the following describe you? patient Strongly inaccurate Inaccurate Slightly inaccurate Slightly accurate Accurate Strongly accurate
FC (Paired)	 Which statement describes you better, and to what extent? I'm interested in learning about the history and politics of other countries. The first thing that I always do in a new place is to make friends. Much more More Slightly more Slightly more More Much more	 Which statement describes you better, and to what extent? honest attentive Much more More Slightly more Slightly more More Much more

The objective of this study

- To investigate whether the length of choice options influences cognitive load and response quality in FC vs Likert formats.

[Hypotheses]

Sentence-type items impose greater cognitive load than adjective-type ones.

1. Response stability is **lower** for sentence-type items.
2. Perceived difficulty is **higher** for sentence-type items.
3. Response time is **longer** for sentence-type items.
4. These effects are more pronounced in the FC format than the Likert format.

Data Collection

Data were collected via crowdsourcing platform (Prolific).

List of conditions and scales used (randomly assigned to one condition)

【item type】 【format】		HEXACO (Ashton & Lee, 2009)	HEXACO adjective (Romano et al., 2023)
		(Short) sentence	Adjective
All 60 items were randomly presented.	Likert (LK)	LK_sentence ($n = 229$)	LK_adjective ($n = 233$)
Create 30 pairs in advance and present them randomly.	Paired Comparison (PC)	PC_sentence ($n = 239$)	PC_adjective ($n = 207$)

Each scale consists of 60 items

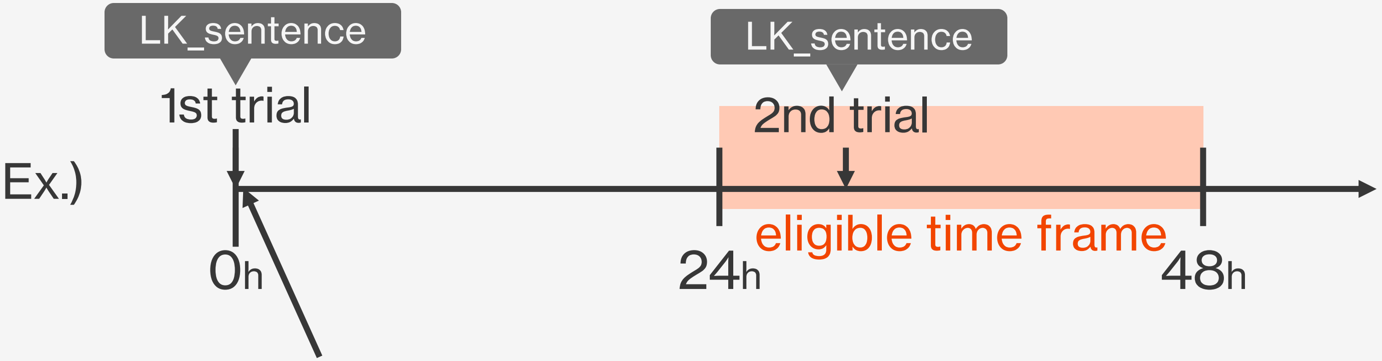
In the preliminary matching of statements, we conducted a genetic algorithm search to minimize the difference in social desirability (collected in a preliminary survey) under the following constraints:

- The number of times factor pairs appear (ensuring each of the 15 pairs appears twice)
- The combination of directions of options (ensuring at least 6 pairs with different directions appear)

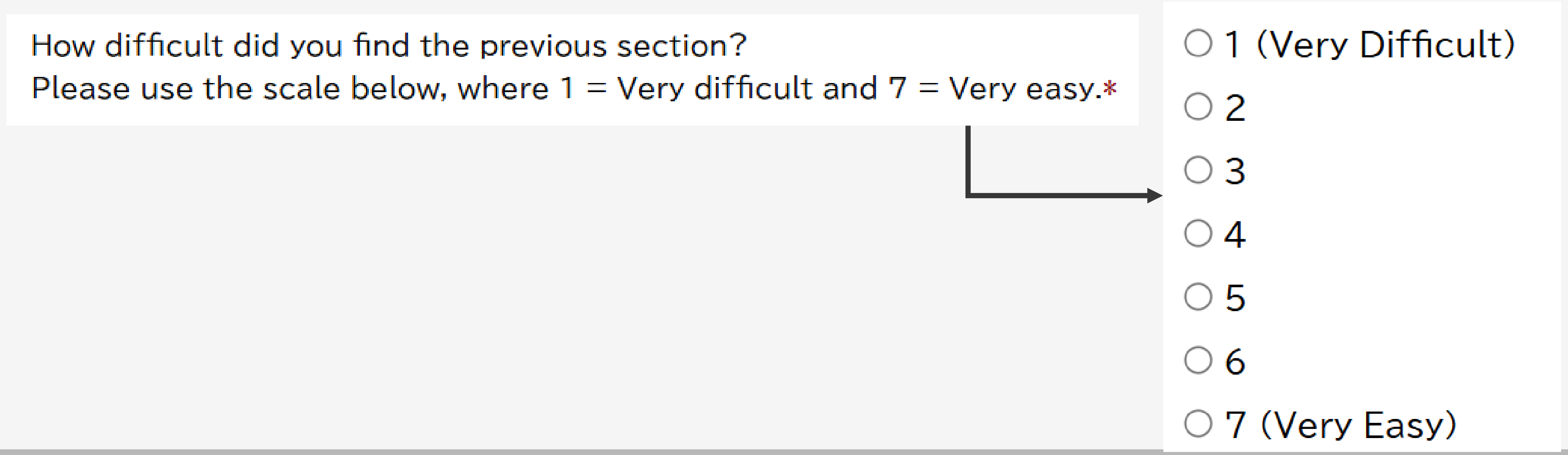
Data collection

Procedure of the data collection

Two sessions: first trial and second trial after 24–48 hours.



Perceived easiness was answered in a 7-point scale after 1st trial

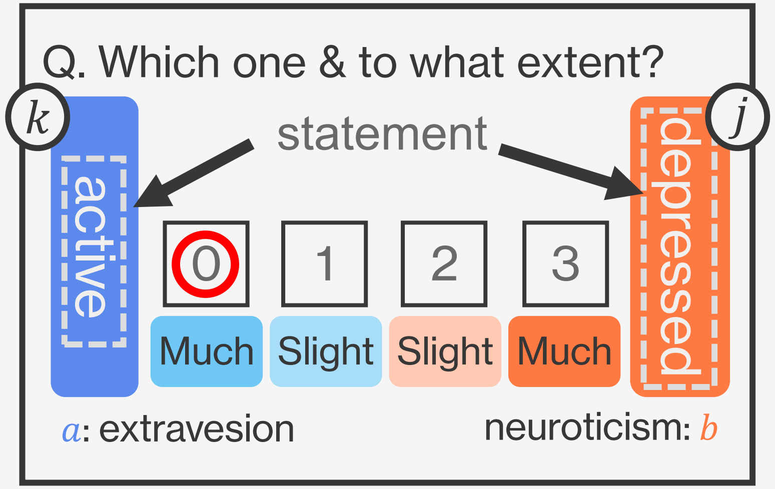

How difficult did you find the previous section?
Please use the scale below, where 1 = Very difficult and 7 = Very easy.*

○ 1 (Very Difficult)
○ 2
○ 3
○ 4
○ 5
○ 6
○ 7 (Very Easy)

Thurstonian IRT model for ordinal response (Brown and Maydeu-Olivares, 2018)

One of the most common models for the FC (Paired Comparison) scale.

Paired Comparison (ordinal)



- Consider a pair of statements (j, k) that reflect different factors (a, b)

$$x_{jk}^{*} = u_j - u_k \qquad x_{jk} = \begin{cases} C-1 & \text{if} \quad x_{jk}^{*} \geq \tau_{C-1} \\ C-2 & \text{if} \quad \tau_{C-1} \geq x_{jk}^{*} \geq \tau_2 \\ \cdots & \\ 1 & \text{if} \quad \tau_2 \geq x_{jk}^{*} \geq \tau_1 \\ 0 & \text{if} \quad x_{jk}^{*} \leq \tau_1 \end{cases}$$

- The latent utility for one statement j is given as:

$$u_j = \mu_j + \beta_j \eta_a + \varepsilon_j \qquad \varepsilon_j \sim N(0, 0.5)$$

fixed to 0.5 for PC scale

- The probability $P(x_{jk} = c|\boldsymbol{\eta})$ is:

Normal ogive model

$$P(x_{jk} \geq c|\boldsymbol{\eta}_i) = \Phi[(\mu_j + \beta_j \eta_a) - (\mu_k + \beta_k \eta_b) - \tau_c]$$

$$P(x_{jk} = c|\boldsymbol{\eta}_i) = P(x_{jk} \geq c|\boldsymbol{\eta}_i) - P(x_{jk} \geq c+1|\boldsymbol{\eta}_i)$$

μ: mean utility of the statement

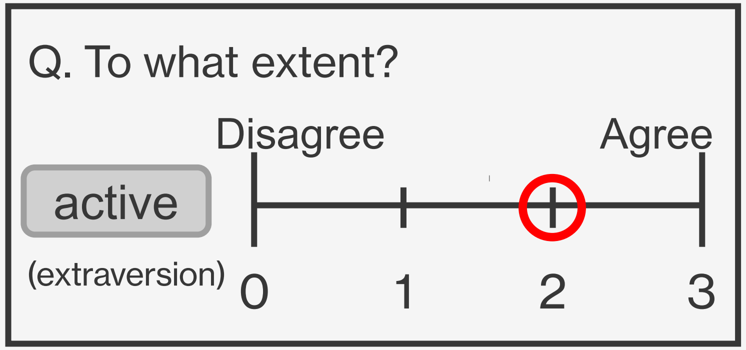
β : factor loading

η : factor score (trait)

Graded response model (Samejima, 1969)

One of the most common models for the Likert scale.

Likert scale



- Consider a statement j that reflect one factor a

$$x_j^* = u_j \qquad x_j = \begin{cases} C-1 & \text{if} \quad x_j^* \geq \tau_{C-1} \\ C-2 & \text{if} \quad \tau_{C-1} \geq x_j^* \geq \tau_2 \\ \quad \cdots & \\ \quad 1 & \text{if} \quad \tau_2 \geq x_j^* \geq \tau_1 \\ \quad 0 & \text{if} \quad x_j^* \leq \tau_1 \end{cases}$$

- The latent utility for the statement j is given as:

$$u_j = \mu_j + \beta_j \eta_a + \varepsilon_j \qquad \varepsilon_j \sim N\left(0, \sigma_j^2\right)$$

- The probability $P\left(x_j = c \middle| \boldsymbol{\eta}\right)$ is:

Normal ogive model

$$P\left(x_j \geq c \middle| \boldsymbol{\eta}_i\right) = \Phi\left[\mu_j + \beta_j \eta_a - \tau_c\right]$$

$$P\left(x_j = c \middle| \boldsymbol{\eta}_i\right) = P\left(x_j \geq c \middle| \boldsymbol{\eta}_i\right) - P\left(x_j \geq c+1 \middle| \boldsymbol{\eta}_i\right)$$

μ: mean utility of the statement

β : factor loading

η : factor score (trait)

Analysis plan

Analysis plan (within a Bayesian framework)

1. Test-retest reliability -> Inspect the posterior distributions of ρ → Explain later
2. Perceived easiness -> Bayesian ANOVA & visually compare distributions
3. Item-wise response-time -> Bayesian ANOVA & visually compare distributions

Bayesian ANOVA

A Bayesian analysis to compute Bayes factor between different models

$$\frac{\boxed{\text{Format}}\quad y = \mu + \beta_{\text{format}} \times \text{format} + e}{\boxed{\text{null}}\quad y = \mu + e} = BF$$

$$\longrightarrow \text{When} \begin{cases} BF > 1, \text{ the main effect of format will be supported.} \\ BF < 1, \text{ the main effect of format will be rejected.} \end{cases}$$

Bayesian estimation

List of estimated parameters and prior distributions

[Respondent i's factor scores]

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\eta}^{(1)} \\ \boldsymbol{\eta}^{(2)} \end{bmatrix} \sim MVN\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}^{(1,1)} & \boldsymbol{\Sigma}^{(1,2)} \\ \boldsymbol{\Sigma}^{(1,2)} & \boldsymbol{\Sigma}^{(2,2)} \end{bmatrix} \right) \qquad \begin{bmatrix} \boldsymbol{\Sigma}^{(1,1)} & \boldsymbol{\Sigma}^{(1,2)} \\ \boldsymbol{\Sigma}^{(1,2)} & \boldsymbol{\Sigma}^{(2,2)} \end{bmatrix} = \boldsymbol{\Sigma} \sim LKJ(1)$$

[mean utility of the statement (or pair)]

$$\left.\begin{array}{ll} \text{(TIRT)} & \alpha\left(= \mu_j - \mu_k\right) \sim normal(0, 5) \\ \text{(GRM)} & \mu_j \sim normal(0,5) \end{array}\right\} \text{with ordered constraint for each statement (or pair)}$$

[factor loadings]

$$\beta_j \sim normal(1,3)$$

Structure of factor score $\boldsymbol{\eta}$

- $\boldsymbol{\eta}$ were estimated as 6*2=12-dimensional parameter

  Same trait was estimated separately (to observe the change between trials)

  (⚠ Item parameters were common between trials)

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\eta}^{(1)} \\ \boldsymbol{\eta}^{(2)} \end{bmatrix} \sim MVN\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}^{(1,1)} & \boldsymbol{\Sigma}^{(1,2)} \\ \boldsymbol{\Sigma}^{(1,2)} & \boldsymbol{\Sigma}^{(2,2)} \end{bmatrix}\right)$$

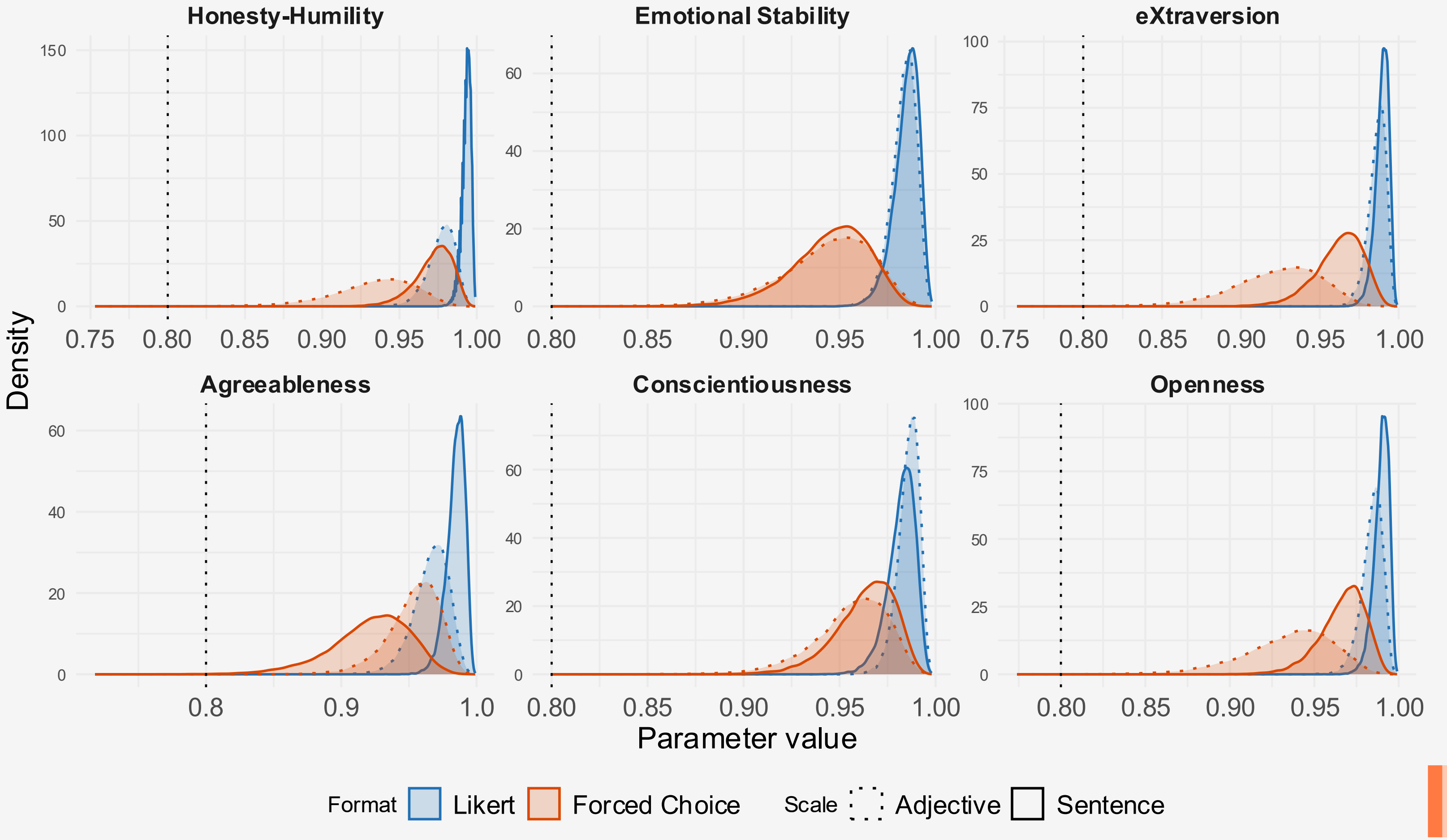
$$\begin{bmatrix} \boldsymbol{\Sigma}^{(1,1)} & \boldsymbol{\Sigma}^{(1,2)} \\ \boldsymbol{\Sigma}^{(1,2)} & \boldsymbol{\Sigma}^{(2,2)} \end{bmatrix} = \boldsymbol{\Sigma} \sim LKJ(1)$$

$$\begin{array}{ll|cccccc|cccccc}
 & & \multicolumn{6}{l}{\text{Trial 1}} & \multicolumn{6}{l}{\text{Trial 2}} \\
 & & \eta_1^{(1)} & \eta_2^{(1)} & \eta_3^{(1)} & \eta_4^{(1)} & \eta_5^{(1)} & \eta_6^{(1)} & \eta_1^{(2)} & \eta_2^{(2)} & \eta_3^{(2)} & \eta_4^{(2)} & \eta_5^{(2)} & \eta_6^{(2)} \\
\hline
\text{Trial 1} & \eta_1^{(1)} & 1 & & & & & & & & & & & \\
 & \eta_2^{(1)} & & 1 & & & & & & & & & & \\
 & \eta_3^{(1)} & & & 1 & & \Sigma^{(1,1)} & & & & & \Sigma^{(1,2)} & & \\
 & \eta_4^{(1)} & & & & 1 & & & & & & & & \\
 & \eta_5^{(1)} & & & & & 1 & & & & & & & \\
 & \eta_6^{(1)} & & & & & & 1 & & & & & & \\
\hline
\text{Trial 2} & \eta_1^{(2)} & \rho_1 & & & & & & 1 & & & & & \\
 & \eta_2^{(2)} & & \rho_2 & & & & & & 1 & & & & \\
 & \eta_3^{(2)} & & & \rho_3 & & \Sigma^{(1,2)} & & & & 1 & \Sigma^{(2,2)} & & \\
 & \eta_4^{(2)} & & & & \rho_4 & & & & & & 1 & & \\
 & \eta_5^{(2)} & & & & & \rho_5 & & & & & & 1 & \\
 & \eta_6^{(2)} & & & & & & \rho_6 & & & & & & 1
\end{array}$$

Diagonal elements in the off-diagonal block matrix can be seen as test-retest reliability (ρ)

Test-retest reliability

Posterior distributions and EAP



format	scale	EAP
FC	Sentence	[0.955, 0.972]
	Adjective	[0.940, 0.956]
LK	Sentence	[0.993, 0.987]
	Adjective	[0.980, 0.986]

[Summary]

- FC reliability is lower than LK but keep $\rho > .8$ (acceptable)
- Sentence-type tends to show higher reliability than adjective

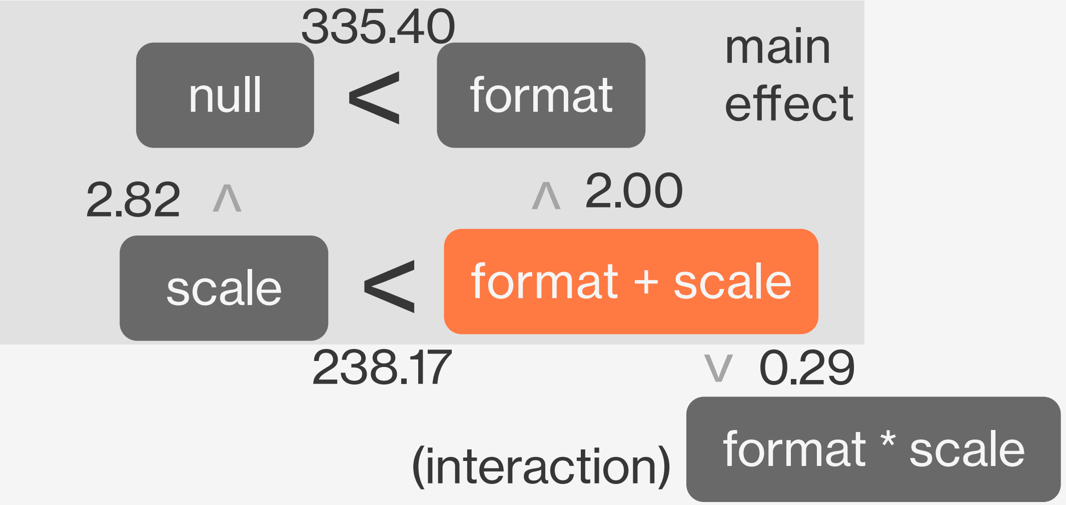
Consists with e.g., Goldberg's (1999) argument

Perceived easiness (7-point scale)

Descriptive statistics

format	scale	mean (SD)
FC	Sentence	5.26 (1.41)
	Adjective	5.60 (1.22)
LK	Sentence	5.71 (1.29)
	Adjective	5.81 (1.29)

Bayesian ANOVA (Bayes factor)



335.40

null < format

main effect

2.82 ∧

∧ 2.00

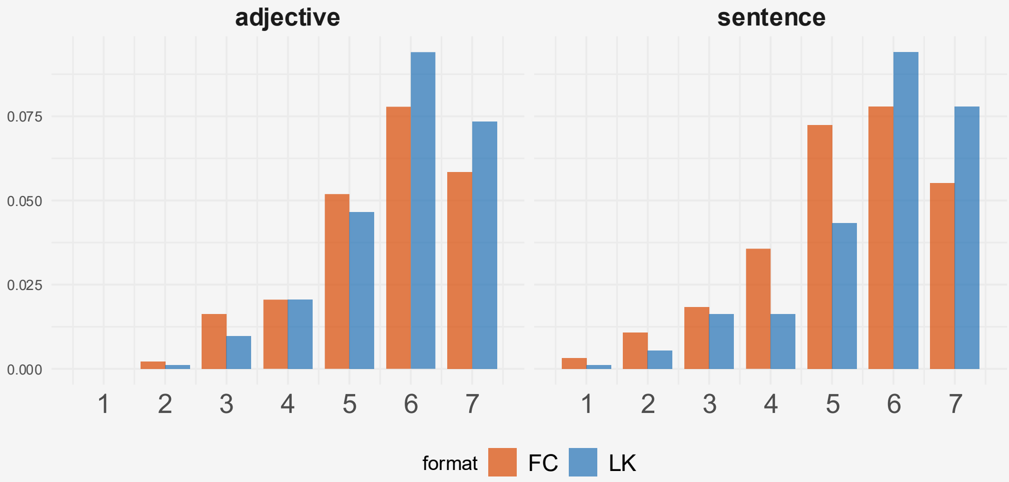
scale < format + scale

238.17

∨ 0.29

(interaction) format * scale

Response distribution (bar plot)



[Summary]

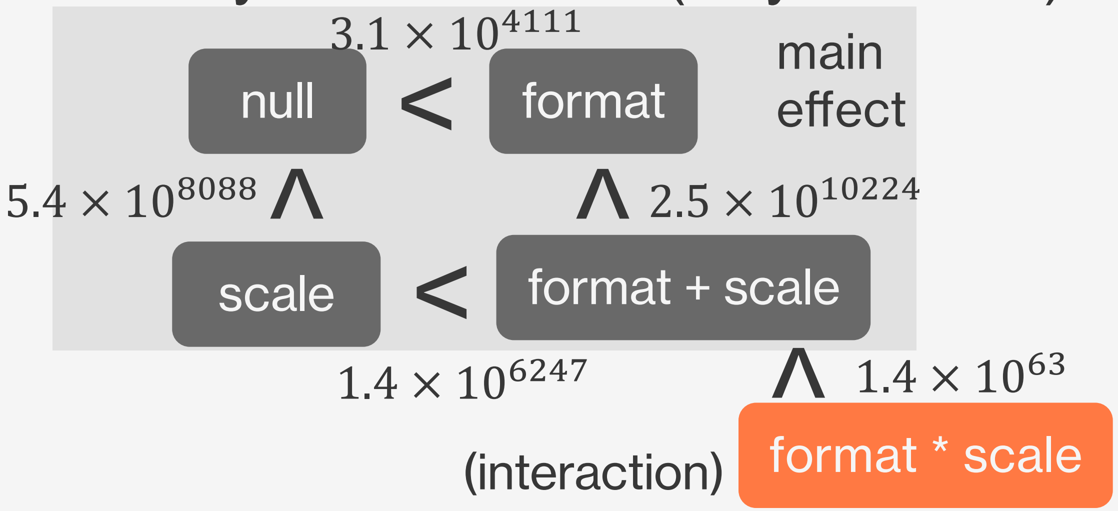
- FC is perceived relatively difficult than LK
- Sentence-type is slightly difficult than Adjective (but the difference is almost negligible)
- No interaction was observed

(log-)Response time

Descriptive statistics (raw RT; sec)

format	scale	median (SD)
FC	Sentence	9.93 (5.87)
	Adjective	4.43 (3.22)
LK	Sentence	5.42 (3.06)
	Adjective	2.66 (2.04)

Bayesian ANOVA (Bayes factor)



3.1×10^{4111}

null < format

main effect

5.4×10^{8088} ∧

∧ 2.5×10^{10224}

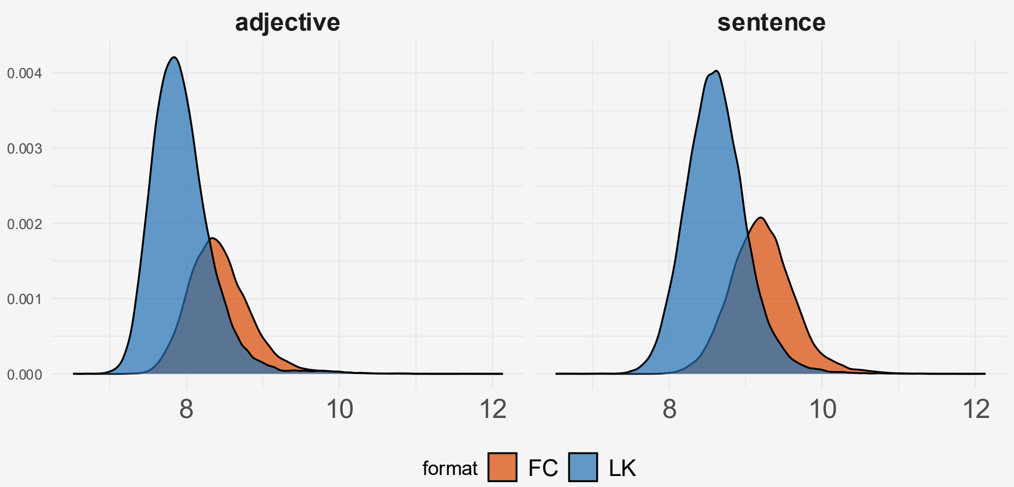
scale < format + scale

1.4×10^{6247}

∧ 1.4×10^{63}

(interaction) format * scale

Response distribution (log-RT)



[Summary]

- FC takes longer than LK (but shorter than 2 times of LK format)
- Sentence-type takes longer than Adjective
  the difference was more salient in FC format

4

Summary and Discussion

Results and future work

Main findings summary

Although item length (Scale) slightly affects test-retest reliability,
the reliability keeps $\rho > 0.8$

→ People can complete the cognitive processes required to answer PC items.

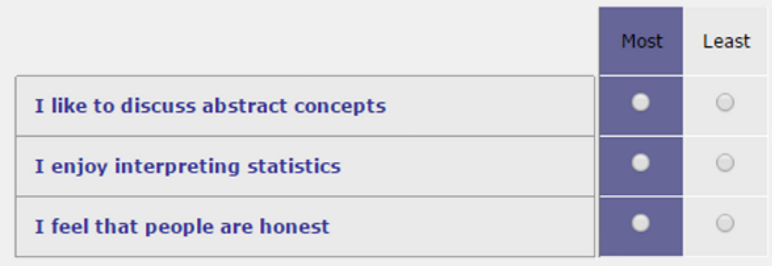
FC seems more difficult than LK, whereas item length does not affect the perception.

Response time per item (block) becomes longer according to the number of words in the item.

Future work

Examine more than 2-alternative FC format

Examine generalizability (with different scales and UIs)



	Most	Least
I like to discuss abstract concepts	●	○
I enjoy interpreting statistics	●	○
I feel that people are honest	●	○

Thank you for your attention!

An investigation of the item length
in the forced-choice psychological measurement

Kyosuke Bunji
Kobe University

Takeshi Sugiyama
Recruit Management Solutions Co., Ltd.

Kensuke Okada
University of Tokyo

✉ bunji@bear.kobe-u.ac.jp

The 13th Conference of the IASC-ARS
Dec 4, 2025

References

Alwin, D. F., & Beattie, B. A. (2016). The KISS principle in survey design: Question length and data quality. *Sociological Methodology*, *46*(1), 121–152. https://doi.org/10.1177/0081175016641714

Ashton, M., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. Journal of Personality Assessment, 91(4), 340–345. https://doi.org/10.1080/00223890902935878

Brown, A., & Maydeu-Olivares, A. (2018). Ordinal factor analysis of graded-preference questionnaire data. Structural Equation Modeling: A Multidisciplinary Journal, 25(4), 516–529. https://doi.org/10.1080/10705511.2017.1392247

Goldberg, L. R. (1999). A broad-bandwidth, public domain personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), Personality psychology in Europe (Vol. 7, pp. 7–28). European Conference on Personality, Tilburg. Tilburg Univ. Press.

Hamby, T., & Ickes, W. (2015). Do the readability and average item length of personality scales affect their reliability? Some meta-analytic answers. Journal of Individual Differences, 36(1), 54–63. https://doi.org/10.1027/1614-0001/a000154

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. Applied Cognitive Psychology, 24(7), 1003–1020. https://doi.org/10.1002/acp.1602

Romano, D., Costantini, G., Richetin, J., & Perugini, M. (2023). The HEXACO adjective scales and its psychometric properties. Assessment, 30(8), 2510–2532. https://doi.org/10.1177/10731911231153833

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika, 34(S1), 1–97. https://doi.org/10.1007/BF03372160

Sass, R., Frick, S., Reips, U.-D., & Wetzel, E. (2020). Taking the test taker’s perspective: Response process and test motivation in multidimensional forced-choice versus rating scale instruments. Assessment, 27(3), 572–584. https://doi.org/10.1177/1073191118762049

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). The psychology of survey response. Cambridge University Press.