日本語版 Semantic Similarity Test の妥当性の検討
Validity of the Japanese version of the Semantic Similarity Test

岡 隆之介[1)]，内海 彰[2)]，楠見 孝[3)]
Ryunosuke Oka, Akira Utsumi, Takashi Kusumi
[1)]三菱電機株式会社，[2)]電気通信大学，[3)]京都大学
Mitsubishi Electric Corporation, The University of Electro-Communications, Kyoto University
Qualia1006@gmail.com

概要

本研究の目的は，呈示された 2 つの単語の類似点を文で回答することで参加者の結晶性知能の一側面を測る課題である，日本語版 Semantic Similarity Test（以下、日本語版 SST）の作成と妥当性を検討することである．予備調査では SST（Stamenković et al., 2019）を参考に，20 項目からなる日本語版 SST の項目と採点基準表を作成した．本調査では，日本語版 SST が収束的妥当性の指標である令和版語彙数推定テスト（藤田・小林, 2022）と弱い正の相関係数を持つことを確認した（$r = .29$）．

キーワード：Semantic Similarity Test, 妥当性, 尺度開発

1. はじめに

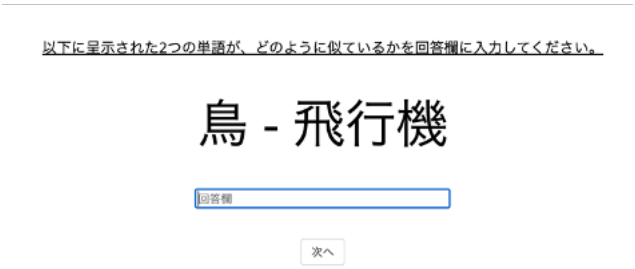
結晶性知能 (crystallized intelligence)は，文化や教育や日々の経験から獲得される知識である[1]．結晶性知能はしばしば，既存の知識では解くことができない問題を解く際に用いられる，流動性知能 (fluid intelligence)と対比して説明される．

臨床場面において，結晶性知能や流動性知能を測定する場合には，Wechsler Adult Intelligence Scale (以下，WAIS)が用いられることが多い．WAIS は言語理解指標やワーキングメモリ指標などからなる，知能を測定するための検査である．なかでも言語理解指標は，単語の意味について回答を求める語義の検査や，2 つの単語の類似点を解凍させる類似の検査によって，結晶性知能を多面的に測定できるという特徴を持つ．一般に，WAIS は，検査のトレーニングを受けた専門家 (例: 臨床検査士)によって測定・評価される．

本研究の大きな目的は，WAIS の類似の検査に替わる，非臨床目的の結晶性知能の測定課題を作成することである．WAIS は臨床目的で使用されるという性質から，問題やその採点基準の詳細が公開できないという制約がある．一方で，結晶性知能は言語理解のさまざまな課題 (例: 比喩理解課題)との関連があるため，結晶性知能を誰でも測定できるような代替課題があることは，認知科学，言語学や心理学に関わるさまざまな領域の研究を進めていく上で重要と考えられる．

本研究では，Stamenkovic et al.[2]によって提案された Semantic Similarity Test (以下，SST)の日本語版 (日本語版 SST)の作成を目指す．SST は，図 1 に示すように，参加者に 2 つの単語を呈示し，呈示された単語間の類似点を自由記述で回答をさせる課題である．

図 1. 日本語版 SST の回答場面



SST は 20 項目からなり，WAIS の「類似の検査」を参考に作られているが，刺激に用いられる項目は全て WAIS と異なるという特徴を持つ．また，それぞれの項目に対して，公開された採点基準表があり，これに基づいて 2 点（2 つの単語の一般的な分類に言及している），1 点（部分的な類似），そして 0 点（類似点が認められない）で採点される．項目，採点基準，および得点の具体例を表 1 に示す．

表 1. 項目，採点基準，および得点の具体例

項目	採点基準	得点
鳥-飛行機	飛ぶ	2
鳥-飛行機	羽がある	1
時間-川	流れる	2
時間-川	連続する	1
時間-川	川の流れのようだ	0
ダイヤモンド-雪片	結晶構造	2
ダイヤモンド-雪片	キラキラしている	1
ダイヤモンド-雪片	脆い (雪片にのみ当てはまる)	0

本研究は予備調査と本調査からなる．予備調査では，SST の英語の単語ペア (e.g., “bird - airplane”)と採点基準 (e.g., 2pts: “can fly”)のそれぞれを日本語に訳し (例:「鳥-飛行機」， 2 点: 「飛ぶ」)，日本語版 SST の原案とする．また，原案の 20 項目が必ずしも日本語との文化的な共通性を持つとは限らないため, 4 項目を追加で予備として準備した．これらの項目について，SST の項目決定の基準に従って信頼性係数 (Chronbach’s alpha)と項目の多様性のバランスで決定した．

本調査では，予備調査で作成した日本語版 SST を用いてその構成概念妥当性を検討する．構成概念妥当性は，令和版語彙数推定テスト[3]と Raven Progressive Matrices Short[4] (以下，RPMS)を用いて行った．令和版語彙数推定テストは，参加者に呈示した単語セットに対して参加者が既知と回答した項目の内容と数に基づいて参加者の語彙数を推定する，ウェブ上で実施される検査である．SST は WAIS-III の語義の検査と正相関 (r = .67; [2])であることが知られているため，語義の検査と同様に参加者の語彙知識を尋ねる令和版語彙数推定テストもまた，今回開発の日本語版 SST と正相関であることが確認されれば，日本語版 SST は収束的妥当性を持つといえるだろう．一方で，RPMS は 12 項目からなる Raven Progressive Matrices[5]の短縮版であり，参加者に 2×8 の行列に漸次的変化する規則性によって排他された抽象的図を呈示し，そこに含まれる空欄部分を最も適切に埋める図形を選択肢の中から選択する課題である．この課題は, 流動性知能を測る標準的な課題の一つである．SST は RPMS と弱い正相関 (r = .31: [2])であることが知られているため, 日本語版 SST と弱い正相関であることが確認されれば，日本語版 SST は弁別的妥当性を持つといえるだろう．

加えて, 探索定な分析として SST と日本語版 Ten Item Personality Inventory (小塩他[6]; 以下, TIPI-J)の相関関係について検討する．TIPI-J は Big Five 性格特性を 10 項目で簡便に測る指標であり，日本語版 SST との関係性を検討することで，結晶性知能に関心のある将来の研究者のための予備的な資料となると考えた．

2. 予備調査

参加者 80 名の参加者 (男性 50 名， 年齢: 27–59 歳， 平均年齢 42. 4 歳) がクラウドワークスで収集された. すべての参加者は匿名で集められた. 予備調査および本調査は京都大学大学院教育学研究科の倫理審査の承認を経て行われた. また, 予備調査と本調査の実施にあたっては，すべての参加者から同意書に対する署名を経た.

刺激 24 個の単語ペアを用いた．項目，採点基準, および得点の具体例を表 1 に示す．20 個の単語ペアとそれぞれの採点基準は SST の筆頭著者 (Dr. Dusan Stamenkovic)の許可を得て著者らによって日本語に翻訳された．また，追加の 4 項目については，単語間類似度について検討した英語の研究[7]の単語ペアから抽出した.

採点基準表 採点基準表は SST と WAIS-III の「類似の検査」を参考に作成された. 採点基準表は全体的な採点基準と項目ごとの採点基準に分かれていた．全体的な採点基準は項目ごとの採点基準がどのように決定されているか (例: 2 点: 「2 つの単語の一般的なカテゴリや普遍的な特徴や関係性について言及されている」，1 点:「部分的な特徴や機能の妥当な類似性について言及されている」，0 点: 「それぞれの単語の特徴について述べらている」,「明らかに誤った回答」)や，採点全体に関する基準がまとめられていた.

手続き データ収集は Qualtrics (http://www.qualtrics.com/; 同意書と参加者の属性情報の収集)と jsPsych (de Leeuw et al.[8]; 日本語版 SST の実施)を用いて行われた．参加者は PC のみから課題に参加可能であった．SST と RPMS をオンラインで実施した Stamenković et al.[9]と同様に，すべてのデータ収集はオンライン上で行われた.

課題は以下の 4 ステップで行われた．第一に，参加者は課題についての説明と同意書への署名を求められた．第二に，参加者は属性情報 (性別，年齢，教育水準)について回答が求められた．教育水準は参加者が現在あるいは最後に在学していた学校で尋ねられ，本調査での分析の際には以下の基準で数値に変換された (1: 中学校, 2: 高校, 3: 短期大学および専門学校, 4: 大学, 5: 大学院)．

第三に，参加者は日本語版 SST に回答した．参加者は，この課題で参加者には単語のペア (具体例として，「椅子 – ソファ」)が呈示されるので，2 つの単語に共通する類似点を回答することが説明された．課題の説明において，WAIS-III と同様に，正解の回答 (例: 「椅子 - ソファ」であれば「家具」)を示すことはしなかった．刺激の呈示順は SST の呈示順に従って 20 項目を最初に呈示し, その後追加の 4 項目を呈示した. SST の項目の提示順序は課題の難易度 (項目の正解率)に基づ

いて，簡単なものから呈示されていた．第四に，参加者の謝金 (300 円)の支払いのための情報を呈示して課題を終了した．本調査は約 16 分で終了した．

結果　日本語版 SST の単語ペアと採点基準表の決定は以下の 5 ステップで行われた．第一に，採点点基準表の原案に基づいて，第一著者がすべての回答を採点した．この際に，項目ごとの採点基準について，採点基準にはないが妥当で複数人からの回答が得られた回答について，新たに採点基準に加えた．第二に，第一著者が改めてすべての回答を採点基準表に基づいて採点した．第三に，第一著者は各参加者の SST の合計点を元に 20 項目を選定した．この時，合計点が著しく低い 1 名の参加者 (5 点/40 点)を以降の分析から除外した．SST と同様に，第一著者と第三著者が信頼性係数と項目のばらつきを考慮して項目を決定した．最終的に選択された 20 項目に基づく Chronbach's alpha は.62 であり，SST の.61 と同程度であった．第四に，第一著者と第三著者で評定者間信頼性を確認するため，項目全体の 11%を抽出してそれぞれ独立に採点した．採点結果の分類一致率を算出したところ，評定者間信頼性は中程度に高かった (Fleiss's kappa = .67; z = 10.1, p < .001)．評定者間で評価が異なる項目については合議の上，不一致を解消した．第五に，第一著者が改めてすべての項目を採点した．

3. 本調査

3.1. 方法

参加者　100 名の参加者 (男性 52 名，年齢：24–58 歳，平均年齢 41.8 歳)がクラウドワークスで収集された．本調査では，日本語版 SST の再検査信頼性を検討(未発表内容)するため，参加者は記名で集められた．

刺激・採点基準表　予備調査で選定された 20 個の単語ペアと採点基準表を用いた．

手続き　データ収集は Qualtrics (同意書，参加者の属性情報，令和版語彙数推定テストの結果，TIPI-J，再検査信頼性のための 2 回目の調査への参加意思確認)と jsPsych (RPMS，日本語版 SST の実施)，そして令和版語彙数推定テスト (https://www.kecl.ntt.co.jp/icl/lirg/resources/goitokusei/vocabulary_test/php/login.php)のウェブサイトを用いて行われた．参加者は PC のみから課題に参加可能であった．

課題は以下の 7 ステップで行われた．第一に，参加者は課題についての説明と同意書への署名を求められた．第二に，参加者は予備調査と同様の属性情報が尋ねられた．第三に，参加者は RPMS に回答した．第四に，参加者は日本語版 SST に回答した．日本語版 SST の手続きは項目数が 20 項目になったことを除いて，予備調査と同様であった．第五に，参加者は令和版語彙数推定テストに回答した．この課題で，参加者はウェブサイトの指示に従って課題を実施し，課題終了後に呈示される参加者の語彙数の推定値を，Qualtrics 上で回答することが求められた．本テストの解析には検査 1 の結果と総合得点を用いた．第六に，参加者は TIPI-J に回答することが求められた．TIPI-J は外向性，協調性，勤勉性，神経症傾向，そして経験に対する開放性 (以下，開放性)の 5 つの下位尺度からなる 10 項目の質問紙である．第七に，参加者の謝金 (500 円)の支払いのための情報と 2 回目の調査の参加意思確認をして，課題を終了した．本調査は約 33 分で終了した．

日本語版 SST の採点　第一に，第一著者が参加者のすべての回答を採点した．第二に，第一著者と第三著者が参加者の回答全体の 22%を個別に採点した．評定者間信頼性は中程度に高かった (Fleiss's kappa = .72; z = 21.2, p < .001)．評定者間で評価が異なる項目については合議の上，不一致を解消した．第三に，第一著者が改めてすべての項目を採点した．

3.2. 結果

日本語版SSTの合計得点(日本語版SST得点)，RPMSの合計得点 (RPMS 得点) 令和版語彙数推定テストの総合得点 (総合語彙数) と第一検査の得点 (第一検査語彙数)，TIPI-J の下位尺度の得点，年齢，そして教育水準の記述統計量と相関係数を表 2 にまとめた．

表2. 各変数の記述統計量と相関係数 (N = 100)

Measure	M	SD	1	2	3	4	5	6	7	8	9	10
1. 日本語版SST得点	28.24	3.57	-									
2. RPMS得点	6.42	2.9	.36 ***	-								
3. 総合語彙数	71694	17415	.29 **	.23 *	-							
4. 第一検査語彙数	73236	18325	.31 **	.27 **	.94 ***	-						
5. 外向性	3.45	1.44	-.20 *	-.16	.00	-.03	-					
6. 協調性	4.96	1.33	-.26 *	-.17 †	.04	.01	.05	-				
7. 勤勉性	4.1	1.29	-.29 **	-.20 *	.08	.07	.17 †	.60 ***	-			
8. 神経症傾向	4.06	1.48	.25 *	.14	-.18 †	-.12	-.46 ***	-.56 ***	-.46 ***	-		
9. 開放性	3.69	1.32	-.03	-.04	-.02	.00	.65 ***	.03	.09	-.32 ***	-	
10. 年齢	41.79	7.72	.03	.03	.14	.20	.05	.08	.14	-.11	.09	-
11. 教育水準	3.35	0.88	.08	.12	.11	.16	-.18	-.10	.08	.21	-.09	-.05

†p <.10, *p < .05, **p < .01, ***p < .001

初めに，日本語版 SST の収束的妥当性を確認するために，日本語版 SST 得点と総合語彙数および第一検査語彙数の相関係数を算出した．結果，日本語版 SST 得点と総合語彙数および第一検査語彙数の相関係数はそれぞれ弱い正の相関係数を示した ($rs = .29, .31, ps < .01$). 続いて，日本語版 SST の弁別的妥当性を確認するため，日本語版 SST 得点と RPMS 得点の相関係数を算出した．結果，日本語版 SST 得点と RPMS 得点は弱い正の相関係数を示した ($r = .36, p < .001$)．最後に，探索的な検討を行った日本語版 SST と TIPI-J の下位尺度の相関係数は，外向性・協調性・勤勉性と弱い負の相関係数を示し ($rs = -.20, -.26, -.29$)，神経症傾向と正の相関係数を示した ($r = .25$).

3.3. 考察

本研究では，日本語版 SST の構成概念妥当性を確認した．収束的妥当性の確認のために日本語版 SST の得点と令和版語彙数推定テストの相関係数を検討したところ，弱い正の相関係数があった。このことから，日本語版 SST は収束的妥当性があると判断した．一方で，日本語版 SST の弁別的妥当性の確認のために RPMS との相関係数を検討したところ，弱い正の相関係数にあり，この相関係数の大きさは日本語版 SST と語彙数の相関係数と同程度であった．ところが，SST の元論文においても，SST と RPMS の相関係数が $r = .31$ であったことから，SST と RPMS の弁別的妥当性は元論文と同程度であると判断した．

日本語版 SST の収束的妥当性が小さかったのは，令和版語彙数推定テストの課題要求が小さかったことに起因していると考えられる．SST の元論文では WAIS-III を用いて SST の収束的妥当性を検討していたが，WAIS-III の語義の検査は単語や絵を参加者に見せてその意味を尋ねる，すなわち再生に基づいて回答を求める課題となっている．一方で，日本語版 SST の収束的妥当性の検討に用いた令和版語彙数推定テストは，参加者に呈示した単語群を参加者が知っているか回答する，すなわち再認に基づいて回答を求める課題となっている．再認は再生よりも課題要求が小さいため，これらのことから，課題の語彙推定に対する感度が小さいため，結果として両検査間の相関係数が小さくなったと考えられる．

4. 総合考察

本研究では，結晶性知能の一側面を検討する，日本語版 Semantic Similarity Test を開発し，その構成概念妥当性を検討した．予備調査から，SST の元論文と同程度の信頼性係数 (Chronbach's alpha = .62)を持つ，日本語版 SST の項目と採点基準表を決定した．本調査では，日本語版 SST が，収束的妥当性の指標である令和版語彙数推定テストと，弁別的妥当性の指標である RPMS と，それぞれ弱い正の相関 ($rs = .29, .36$)であることを確認した．日本語版 SST が SST と比して収束的妥当性が弱い結果となったが，このことは先行研究と本研究の語彙数の推定方法の違いで生じていると考察した．以上の結果から，日本語版 SST は収束的妥当性がやや小さいものの，SST の元論文と同程度の収束的妥当性を持ち，結晶性知能の検査として有効と判断した．

今後は，本研究で得られた項目を論文と合わせて公開し[10]，結晶性知能と関連の深い認知課題 (例: 比喩理解課題)と合わせて日本語版 SST を用いて，結晶性知能と認知機能の関係を検討していく．

文献

[1] Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. Cambridge, UK: Cambridge University Press.

[2] Stamenković, D., Ichien, N., & Holyoak, K. J. (2019). Metaphor comprehension: An individual-differences approach. *Journal of Memory and Language*, *105*, 108-118.

[3] 藤田早苗・小林哲生 (2022). 令和版単語親密度に基づく大規模語彙数推定調査 ~Web 公開版の利用ログ分析~, 2022 年度人工知能学会全国大会.

[4] Arthur, W., Tubre, T. C., Paul, D. S., & Sanchez-Ku, M. L. (1999). College-sample psychometric and normative data on a short form of the raven advanced progressive matrices test. *Journal of Psychoeducational Assessment*, *17(4)*, 354–361.

[5] Raven, J. C., Raven, J., & Court, J. H. (1994). *A manual for Raven's Progressive Matrices and Vocabulary Scales*. London, UK: H. K. Lewis.

[6] 小塩真司・阿部晋吾・ピノ カトローニ (2012). 日本語版 Ten Item Personality Inventory (TIPI-J) 作成の試み. パーソナリティ研究, *21(1)*, 40-52.

[7] Xu, X. (2010). Interpreting metaphorical statements. *Journal of Pragmatics*, 42(6), 1622-1636.

[8] de Leeuw, J.R., Gilbert, R.A., & Luchterhandt, B. (2023). jsPsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, *8(85)*, 5351-5354.

[9] Stamenković, D., Milenković, K., Ichien, N., & Holyoak, K. J. (2023). An individual-differences approach to poetic metaphor: Impact of aptness and familiarity. *Metaphor and Symbol*, *38(2)*, 149-161.

[10] Oka, R., Utsumi, A., & Kusumi, T. (under review). Validity and the reliability of the Japanese version of the Semantic Similarity Test.