二重課題は言語モデルの合理的な言語理解ストラテジーを促進する

江村 玲 [1,2,3]　菅原 朔 [2]
[1] 東北大学文学研究科　[2] 国立情報学研究所　[3] 日本学術振興会特別研究員 DC
rei.emura.r4@dc.tohoku.ac.jp　saku@nii.ac.jp

概要

計算に使うメモリ資源を制限すると，言語モデルは人間の読み時間をより正確に予測する．本研究は，この知見は言語理解ストラテジーにも当てはまるのか調べた．具体的には，GPT-4o は，人間と同じように，メモリ資源が制限されると，非妥当な文を妥当な意味で理解するという合理的な言語理解ストラテジーを使用しやすくなるか検証した．メモリ資源の制約のために，“The 2 cocktail + blended 3 =...” のように，計算問題と言語理解を同時に行う二重課題を設計して実行した．結果，このようなストラテジーの変化を観察した．これは，合理的な言語理解ストラテジーへの転換はメモリ資源の制限が原因の一つであることを支持する．

1　はじめに

1.1　メモリ資源の制限は人間らしい言語理解をもたらす

計算心理言語学者は，言語モデルと人間を比較することで，「人間らしい」言語理解とはどのようなものか，そしてどんな性質が人間を人間たらしめているのかを研究してきた．計算に使うメモリ資源（人間でワーキングメモリ[1)]に当たる）を制限すると，言語モデルがより人間の振る舞いに近くなることから，「人間らしい」言語理解を作り出す性質としてメモリ資源の制限が挙げられている．

多く研究されているものは，人間の読み時間を予測する Surprisal [2, 3] を言語モデルから計算する方法である．具体的には，次の語を予測するための文脈に制限をつけたり，短くしたり，また self-attention head を一つにしたりすることで Surprisal を計算すると，人間の読み時間をより正確に予測した [4, 5, 6].

1)　ワーキングメモリとは，処理に使われるため，引き出し可能な状態で保管されるメモリである [1].

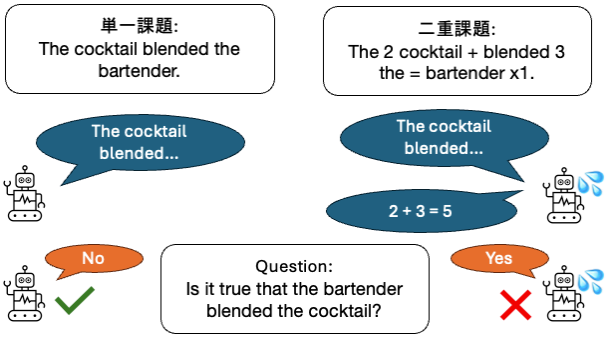

図 1　仮説と課題の概要図．言語モデルが計算問題と言語理解を同時に行う二重課題に取り組むことで，人間と同じように言語モデルも，非妥当な文を理解するときに文法よりも妥当性を優先するのか確かめる．

したがって，計算に使うメモリ資源の制限は，言語理解の処理負荷（つまり，文中の任意の語の処理の難しさ）に関して人間らしさをもたらす．

1.2　理解ストラテジーはメモリ資源によって変わるのか

しかし，処理負荷だけではなく言語理解の方略に関しても，メモリ資源を制限することで言語モデルは人間らしくなるかは，まだ明らかになっていない．以下，詳細を説明する．

人間は，合理的な言語理解ストラテジーを用いることがある．このストラテジーは，(2) のような文字通りでは非妥当な文を読んだとき，語の並びを入れ替えて，(1) のような妥当な意味（一般常識と合致する内容）として解釈することを指す．つまり，文法情報よりも妥当性の情報を優先するストラテジーだと言える．人間においてこのストラテジーは，二重課題下など，認知負荷が高い場合に使用されやすくなることが観察されている [7, 8, 9, 10].

(1)　妥当な文：*The bartender blended the cocktail.*

(2)　非妥当な文：*The cocktail blended the bartender.*

では，言語モデルも，人間と同じように合理的な

言語理解ストラテジーへの転換を行うのだろうか．たしかに，言語モデルでも表面的に文を理解する現象は多く観察されている [11, 12]．また人間と同じように，文長が長くなると，非妥当的な文を妥当的な文として理解する傾向が高まった [13]．しかし，文長が長くなると妥当な文に対する正確性も下がるため，文の長さがストラテジーの転換をもたらしたのかは確かではない．なお，self-attention head の数は，このストラテジー転換には影響しなかったと報告されている [13]．

そこで本研究は，課題でメモリ資源を操作することを試みる．具体的には，言語理解と計算問題の二つの課題を実行する二重課題を実施する．

1.3 研究目的

本研究の目的は，**言語モデルは計算資源が制限された状況では，人間のように合理的な言語理解ストラテジーを示すか**，二重課題を用いて調べることである．具体的な仮説として，「**単一課題よりも二重課題おいて，言語モデルは非妥当な文を与えられたとき，単語を入れ替えて妥当的な解釈として理解する傾向が高くなる**」を掲げる（図 1 参照）．

本研究が指す二重課題とは，二つの処理（計算問題+言語理解）と記憶（文の内容）のバランスを保つ必要があるものを指す．二重課題の設計においては，人間のワーキングメモリ容量を測定する課題を参考にした．なお，自然言語処理研究において，マルチタスク下では正確性が下がることがすでに観察されている [14, 15, 16]．しかし，これらの研究におけるマルチタスクとは複数の課題が独立して順に並んだものを指し，本研究のようにプロンプトの中に複数の課題が交錯しているものはまだ試みられていない．

本研究の実験では，妥当な文と非妥当な文を文脈として，単一課題と二重課題の質問応答課題を実行した．そして言語モデルは，二重課題では単一課題に比べて，妥当な文に対する非妥当な文の正答率の低下が著しくなるのか確かめた．

2 実験方法

2.1 課題

我々は，以下の 3 種類の質問応答課題を実行した．二重課題は，人間のワーキングメモリの容量を計測するための Operation Span Task [17] を参考に，言語モデルの仕様（文字列の羅列の方が適している）に合わせて作成した．ノイズのある単一課題は，単純に計算式というノイズのある文脈文が影響するのか，それとも二重課題という追加の課題が影響するのかを確認するために行なった．

- **単一課題**：計算問題が埋め込まれていない文脈文をインプットする．その後文脈文に関する質問に答える．
- **ノイズのある単一課題**：計算問題付きの文脈文を，計算問題を無視しながらインプットする．その後文脈文に関する質問に答える．
- **二重課題**：計算問題付きの文脈文を，計算問題に回答しながらインプットする．その後文脈文に関する質問に答える．

言語モデルには GPT-4o を用いた [18]．GPT-4o は，Trasnfomer [19] を基にしており，GPT 系列の中で現在最も強力なモデルの一つである．GPT-4o の system message に課題のプロンプト（付録 A を参照）を入力し，user message に文と質問を入力した．

モデルとプロンプトの選定方法は，(1) 単一課題で非妥当条件の正答率が 70%以上であること，(2) 二重課題における計算問題の正答率が 95%以上であることという条件を満たすことである．(1) の条件で単一課題で非妥当条件の問題を解けていることを確認した．(2) の条件では実際に言語モデルが二重課題の計算問題を行なっていることを確認した．GPT-3.5-turbo[2] でも実行したが，(1) と (2) の条件を満たさなかったため選定しなかった．

2.2 データセット

データセットは，GELP データセット [13] の一部を用いた．[3] このデータセットには，文と質問のペアが含まれている．文には，(1) と (2) のように，二つの妥当性のタイプ（妥当/非妥当）がある．主語に生物，目的語に無生物をとる動詞に対して，主語に生物，目的語に無生物を入れた文を妥当な文とし，主語に無生物，目的語に生物を入れた文を非妥当である文とした．また，異なる 8 つの文構造が含まれているが，これらは非妥当条件で人間が誤解を生じやすいことが確認されたものである．この基本の文に，文を長く複雑にするために，2 つの命題が接続詞で接続されている ((3) を参照) ．全部の刺激

2) https://platform.openai.com/docs/models
3) データセットは https://github.com/nii-cl/gelp から入手し，Memory Load が High である条件の文を用いた．

は，2560 ペア（2 つの妥当性タイプ * 8 つの文構造 * 160）である．

質問は，基本の文に関して Yes/No で答える二択問題である．質問の文構造はすべて (5) のように，*Is it true that* に続けて基本の文が置かれている．また，答えが No である質問はすべて，主語と目的語が入れ替わったものである．正答が Yes と No の質問が半分ずつ含まれている．

さらに，二重課題を実施するために，(4) のような計算問題付きの文をこのデータセットに追加した．ランダムに計算問題を生成し，=のあとは x1, x2, x3... という文字列を追加した．この計算問題を一語ずつ交互に文に埋め込んだ．計算問題の途中で文の末尾に到達した場合，残りの計算問題は加えなかった．計算問題は，1 桁数字を 2 つ足す問題（**1 桁–2 足し算**），1 桁数字を 3 つ足す問題（**1 桁–3 足し算**），3 桁数字を 2 つ足す問題（**3 桁–2 足し算**），3 桁数字を 3 つ足す問題（**3 桁–3 足し算**）の 4 種類がある．

(3) 文：*The cocktail blended the bartender and the intruder cited the patent after the neurologist baffled the hippie.*

(4) 計算問題付き文：*The 7 cocktail + blended 7 the + bartender 3 and = the x1 intruder 7 cited + the 1 patent + after 9 the = neurologist x2 baffled 3 the + hippie.*

(5) 質問：*Is it true that the bartender blended the cocktail?*

3 実験結果・考察

データの前処理として，回答を抽出できなかったデータを分析から外した．[4] その結果，質問応答のデータは 3.7%，計算問題のデータは 5.6%削除した．

質問応答の回答結果は，課題と妥当性ごとに見ると，表 1 のようになった．なお，二重課題における計算問題の正答率の平均値は，すべての計算問題で 99.5%以上だった．[5]

3.1 課題と妥当性の主効果

質問応答課題のデータを計算問題別に，二項分布の一般化線形混合モデルを用いて分析した．[6] 一般化線形混合モデルには，固定効果として課題（二重課題/ノイズのある単一課題/単一課題）と妥当性（妥当/非妥当）が含まれる．ランダム効果としてアイテムが含まれる．課題要因は 3 つの水準があるため，二重課題を基準とした．

結果，すべての計算問題において，課題（単一課題）と妥当性の主効果が $\alpha = 0.05$ で有意となり，単一課題に比べ二重課題の正答率が低く，妥当な文に比べ非妥当な文の正答率が低かった．また，1 桁–3 足し算と 3 桁–3 足し算以外で，ノイズのある単一課題に比べ二重課題の正答率が有意に低かった．

これらのデータは，課題の二重性に関わらず，言語モデルは文法情報と妥当性の情報の両方から文を理解し，また妥当性に関わらず，二重課題で文理解の正確性が低くなることを示した．これらは，言語モデルは表面的な言語理解をするという先行研究 [11, 12] や，複数のタスクを行うことで正確性が落ちるという先行研究 [14] を再現したものである．

3.2 課題ごとの妥当性の効果

次に，課題と妥当性の交互作用のうち，どの課題が妥当性の効果に寄与したかを調べるため，事後推定平均を求め，課題の条件ごとに妥当性のオッズ比 *odds*(妥当条件)/*odds*(非妥当条件) を調べた．[7]

表 2 に示す通り，1 桁–2 足し算以外のすべての計算問題で，二重課題でのオッズ比が単一課題やノイズのある単一課題でのものよりも高くなった．このことは，妥当性が正答率に与える影響が，二重課題の時に最も大きくなったことを示す．したがって，**言語モデルは，二重課題下では単一課題下に比べて，合理的な言語理解ストラテジーを使用する傾向が高くなるとが示唆された**．なお，1 桁–2 足し算の条件でこの傾向が見られなかったのは，計算問題の負荷が小さすぎるためであると推測する．

また，文構造によって妥当性の効果の違いがあるかを調べるために．文構造と課題，妥当性ごとに平均正答率を算出した（データは付録 B を参照）．結果，2 種類の二重目的語構文は，妥当性の効果が二重課題で大きくなることは観察されなかった．これは，GPT-2 を用いた研究と同様に，妥当な文の正答率が他の構文よりも低いことが理由だと考える [13]．

さらに，正答によって妥当性の効果の違いがあるかを調べるために，正答と課題，妥当性ごとに平均正答率を算出した（詳細は付録 C を参照）．正答が

4) データの分析はすべて R version 4.3.2 [20] を用いた．
5) すべての結果の詳細は https://github.com/reiemura/rational_llm_nlp2025 を参照のこと．
6) lme4 パッケージ [21] を用いた．
7) emmeans パッケージ [22] を用いた．

表 1　それぞれの計算問題・課題・文の妥当性における正答率の平均値.

計算問題	課題	妥当性		
		妥当	非妥当	Δ 妥当 – 非妥当
なし	単一課題	0.91	0.77	0.14
1 桁–2 足し算	二重課題	0.87	0.57	0.30
	ノイズのある単一課題	0.91	0.64	0.26
1 桁–3 足し算	二重課題	0.87	0.57	0.30
	ノイズのある単一課題	0.89	0.66	0.23
3 桁–2 足し算	二重課題	0.87	0.56	0.31
	ノイズのある単一課題	0.90	0.64	0.25
3 桁–3 足し算	二重課題	0.87	0.58	0.29
	ノイズのある単一課題	0.88	0.63	0.25

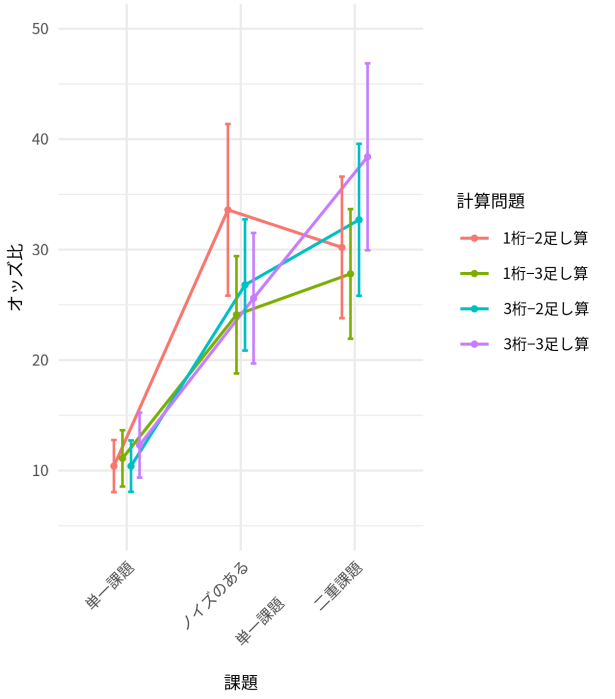

図 2　妥当性のオッズ比 *odds*(妥当条件)/*odds*(非妥当条件) の平均値．エラーバーは標準誤差を示す．オッズ比が大きいことは，その課題が妥当性が正答率に与える効果に大きく寄与していることを示す.

No であるときは妥当性の効果が二重課題で大きくなることが観察されたが，正答が Yes であるときは観察されなかった.

3.3　メモリ資源の制限は合理的な言語理解ストラテジーを促進する

上記の結果は，計算問題や文構造，正答により差があるものの，言語モデルは二重課題下で合理的な言語理解ストラテジーを促進するという傾向を示した．これは，文法情報よりも妥当情報を優先してしまうエラーの原因の一つは，二重課題のように言語理解に割く資源が少なくなることだと示唆する.

最後に，本研究の心理言語学分野への貢献について記す．人間でも合理的な言語理解ストラテジーへの転換が観察されたが，それらは二重課題下のほか，文構造が複雑な場合や第二言語話者の場合など，認知負荷が高い状況である [7, 8, 9, 10]．このことから，このストラテジーの転換の動機は，計算に使うメモリ資源，つまりワーキングメモリを節約するためだと考察された [23, 9] .[8)] 本研究は新たに，人間より大きなメモリ容量を持つと思われる言語モデルにおいても合理的な言語理解ストラテジーへの転換を観察し，人間についての仮説が同様に支持される可能性を示した.

4　おわりに

本研究は，GPT-4o に対して質問応答課題と計算問題の二重課題を実行し，このようなメモリ資源が制限された条件では，文法情報より妥当性情報を優先する傾向が高くなることを観察した.

本研究の限界の一つは，妥当性の効果への影響が，二重課題とノイズのある単一課題で顕著な違いを観察できなかった点である．1 桁の計算より 3 桁の計算で効果が大きくなったため，今後はさらに負荷の高い二重課題を行い，さらに二重課題の効果が大きくなるのか確認すべきである．また本研究は言語モデルのみを観察したので，今後は比較のため人間を対象とした実験を実施することが必要である.

8)　[23] は，[8] や [9] とは理論とは別の理論を提示している.

謝辞

本研究は JSPS 科研費 JP23KJ0199, JST 創発的研究支援事業 JPMJFR232R の助成を受けたものです.

参考文献

[1] Alan D Baddeley, Neil Thomson, and Mary Buchanan. Word length and the structure of short-term memory. **Journal of verbal learning and verbal behavior**, Vol. 14, No. 6, pp. 575–589, 1975.

[2] J Hale. A probabilistic earley parser as a psycholinguistic model. In **Proceedings of the Second Meeting of the North American Chapter of the Asssociation for Computational Linguistics**, pp. 159–166, 2001.

[3] Roger Levy. Expectation-based syntactic comprehension. **Cognition**, Vol. 106, No. 3, pp. 1126–1177, 2008.

[4] Michael Hahn, Richard Futrell, Roger Levy, and Edward Gibson. A resource-rational model of human processing of recursive linguistic structure. **Proceedings of the National Academy of Sciences**, Vol. 119, No. 43, p. e2122602119, 2022.

[5] Tatsuki Kuribayashi, Yohei Oseki, Ana Brassard, and Kentaro Inui. Context limitations make neural language models more human-like. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 10421–10436, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[6] William Timkey and Tal Linzen. A language model with limited memory capacity captures interference in human sentence processing. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 8705–8720, Singapore, December 2023. Association for Computational Linguistics.

[7] Richard Futrell and Gibson Edward. L2 processing as noisy channel language comprehension. **Bilingualism: Language and Cognition**, Vol. 20, No. 4, pp. 683–684, 2017.

[8] Edward Gibson, Steven T. Piantadosi, Kimberly Brink, Leon Bergen, Eunice Lim, and Rebecca Saxe. A noisy-channel account of crosslinguistic word-order variation. **Psychological Science**, Vol. 24, No. 7, pp. 1079–1088, 2013. PMID: 23649563.

[9] Evelina Fedorenko Leon Bergen Edward Gibson, Chaleece Sandberg and Swathi Kiran. A rational inference approach to aphasic language comprehension. **Aphasiology**, Vol. 30, No. 11, pp. 1341–1360, 2016.

[10] Corianne Rogalsky, William Matchin, and Gregory Hickok. Broca's area, sentence comprehension, and working memory: an fmri study. **Frontiers in human neuroscience**, Vol. 2, p. 327, 2008.

[11] Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. Shortcut learning of large language models in natural language understanding, 2023.

[12] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. **Nature Machine Intelligence**, Vol. 2, No. 11, pp. 665–673, 2020.

[13] Daiki Asami and Saku Sugawara. What makes language models good-enough? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 15453–15467, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[14] Zhoujun Cheng, Jungo Kasai, and Tao Yu. Batch prompting: Efficient inference with large language model apis, 2023.

[15] Guijin Son, SangWon Baek, Sangdae Nam, Ilgyun Jeong, and Seungone Kim. Multi-task inference: Can large language models follow multiple instructions at once? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 5606–5627, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[16] Zhengxiang Wang, Jordan Kodner, and Owen Rambow. Exploring the zero-shot capabilities of llms handling multiple problems at once, 2024.

[17] Andrew RA Conway, Michael J Kane, Michael F Bunting, D Zach Hambrick, Oliver Wilhelm, and Randall W Engle. Working memory span tasks: A methodological review and user's guide. **Psychonomic bulletin & review**, Vol. 12, pp. 769–786, 2005.

[18] OpenAI. GPT-4 technical report, 2024.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Advances in Neural Information Processing Systems**, 2017.

[20] R Core Team. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, 2023.

[21] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. **arXiv preprint arXiv:1406.5823**, 2014.

[22] Russell V. Lenth. **emmeans: Estimated Marginal Means, aka Least-Squares Means**, 2023. R package version 1.8.9.

[23] Kiel Christianson, Carrick C Williams, Rose T Zacks, and Fernanda Ferreira. Younger and older adults'" good-enough" interpretations of garden-path sentences. **Discourse processes**, Vol. 42, No. 2, pp. 205–238, 2006.

A　プロンプトの内容

GPT-4o の system message へのプロンプトは以下を用いた．なお，ノイズのある単一課題と二重課題において 3 つの足し算の条件では，下記のように 2 つではなく，3 つの足し算の例に変えている．

単一課題：Please read the sentence. Once the sentence ends, answer the following question. The question relates to the sentence and requires a yes or no response. Return the question answers as [Yes] or [No].

ノイズのある単一課題：The sentence combines a math problem and a text alternating one word at a time. The math problem is split and placed after each word in the sentence, such as 2 + 8 = x1. Please read the sentence while ignoring the math problems. Once the sentence ends, answer the following question. The question relates to the sentence and requires a yes or no response. For example, in the case below, ignore the letters 2, +, 8, =, x1, 1, +, 2, =, x2, 5, +, 6, =, x3, 6, and +. sentence = The 2 quick + brown 8 fox = jumps x1 over 1 the + lazy 2 dog = on x2 a 5 sunny + day 6 in = the x3 beautiful 6 green + park. Return the question answers as [Yes] or [No].

二重課題：The sentence combines a math problem and a text alternating one word at a time. The math problem is split and placed after each word in the sentence, such as 2 + 8 = x1. Please read the sentence while accurately calculating the math problems. When x1, x2, x3... appear, output the answer to the math problem carefully. Once the sentence ends, answer the following question. The question relates to the sentence and requires a yes or no response. However, prioritize the quality of solving the math problems over answering the sentence. Ensure all math problems are solved correctly, with each one being an addition of three numbers. For example, in the case below, output the answer to 2 + 8 at x1. Output the answer to 1 + 2 at x2. sentence = The 2 quick + brown 8 fox = jumps x1 over 1 the + lazy 2 dog = on x2 a 5 sunny + day 6 in = the x3 beautiful 6 green + park. Each math problem is always an addition of two numbers like this. Return the math problems and their answers as tuples, e.g., (x1, 2 + 8, 10), (x2, 1 + 2, 3). Return the question answers as [Yes] or [No].

B　構造ごとの妥当性の効果

データセットは以下の 8 種類の文構造を含む．

- 他動詞文：The bartender blended the cocktail.
- 受動文：The cocktail was blended by the bartender.
- 与格文：The employee gave the book to the sister.
- 二重目的語文：The employee gave the sister the book.
- 受益文：The hostess served the tea for the guest.
- 受益二重目的語文：The hostess served the guest the tea.
- 経験者の主語：The designer favored the style.
- 経験者の目的語：The equation confused the mathematician.

表 2　それぞれの構造・課題・妥当性における正答率の平均値．

構造	課題	妥当性		
		妥当	非妥当	Δ妥当 – 非妥当
他動詞文	単一課題	1.00	0.91	0.09
	ノイズのある単一課題	1.00	0.64	0.36
	二重課題	1.00	0.53	0.47
受動文	単一課題	1.00	0.90	0.10
	ノイズのある単一課題	0.98	0.68	0.30
	二重課題	0.97	0.61	0.36
与格	単一課題	0.99	0.70	0.29
	ノイズのある単一課題	0.99	0.59	0.40
	二重課題	0.96	0.54	0.42
二重目的語文	単一課題	0.64	0.48	0.16
	ノイズのある単一課題	0.64	0.56	0.08
	二重課題	0.56	0.49	0.06
受益文	単一課題	0.93	0.84	0.10
	ノイズのある単一課題	0.95	0.61	0.33
	二重課題	0.95	0.52	0.42
受益二重目的語文	単一課題	0.76	0.57	0.19
	ノイズのある単一課題	0.64	0.66	-0.02
	二重課題	0.58	0.59	0.00
経験者の主語	単一課題	0.99	0.89	0.11
	ノイズのある単一課題	0.99	0.75	0.24
	二重課題	0.98	0.65	0.33
経験者の目的語	単一課題	0.98	0.88	0.11
	ノイズのある単一課題	0.99	0.67	0.32
	二重課題	0.96	0.64	0.32

C　正答ごとの妥当性の効果

表 3　それぞれの計算問題・課題・妥当性における正答率の平均値．

正答	課題	妥当性		
		妥当	非妥当	Δ妥当 – 非妥当
Yes	単一課題	0.96	0.83	0.13
	ノイズのある単一課題	0.96	0.88	0.08
	二重課題	0.97	0.83	0.14
No	単一課題	0.86	0.71	0.15
	ノイズのある単一課題	0.83	0.41	0.42
	二重課題	0.77	0.31	0.46