2023-12-19

M1　櫛田

Tal Waltzer, Riley L. Cox, and Gail D. Heyman. (2023).

Testing the Ability of Teachers and Students to Differentiate between Essays Generated by ChatGPT and High School Students.

Human Behavior and Emerging Technologies, 2023, 1923981.
https://doi.org/10.1155/2023/1923981

1. Introduction

◆ 人工知能（AI）は人々の生活を根本的に変える可能性を秘めている。

- 学者たちは、仕事を遂行するための人々と AI の協力(Lopes et al., 2022)や、AI が雇用の安全に脅威を与える可能性(Aljanabi & ChatGPT, 2023; Susskind, 2020; Weidinger et al., 2021)、医療における AI の役割(Baxter et al., 2019; Lin et al., 2021)、公衆衛生問題に対する人々の意識に AI が与える影響(Biswas, 2023)などを検討してきた。

- AI の教育への影響も疑問であり、さまざまな状況でどの教育アプローチが最も効果的かの特定(Luckin et al., 2016)や、教育における差別など(Akgun & Greenhow, 2022)では、社会問題を悪化させる可能性もある。

- 教育におけるもう 1 つの懸念は、学問の誠実さを損なう可能性があるということである。
  - ChatGPT は、広範囲のプロンプトに対する応答を生成できるため、人間が作成した文章と区別するのが難しいと思われる。学生は学業をカンニングするために ChatGPT を使用する誘惑に駆られる可能性があり(Susnjak, 2022; Cotton et al., 2023; Mitchell, 2022; Stokel-Walker, 2022; Zhuo, 2023)、一部の学校ではこの理由から既に使用を禁止している(Rosenzweig-Ziff, 2023)。

◆ ChatGPT や他の AI が学術的完全性に悪影響を与えることについての懸念が広く広まっているにもかかわらず、この問題についての実証データはほとんどない。

- 学術的誠実性に対する実際の脅威がどの程度存在するかを評価する必要性がある。
- 不正行為に関連した ChatGPT に対する人々の態度、および教育における AI の使用についてのより一般的な態度に関するデータも重要（Celik, 2023; Chocarro et al., 2023)。

- ✧ AI に対する態度は、AI が教室でどれほど有用な役割を果たすかに影響を与える可能性がある(Zhai et al.,2021)。
- ✧ また、そのテクノロジーがどのように受け入れられ、実装されるかに影響を及ぼす(Kelly et al., 2022)。

◆ 本研究では、高校における教師と生徒を評価することによって、文献におけるこれらのギャップに対処しようとしている。

2. Literature Review

◆ ChatGPT のリリース前から、いくつかの研究が ChatGPT の前身を調査した。
- ➢ ある研究では、GPT-2 によって生成された詩と人間が書いた詩を区別するように求められた課題で、参加者は悪い成績を収めた(Kobis & Mossink, 2021)。
  - ✧ GPT-2 生成セットから最良の詩を厳選した場合、パフォーマンスは偶然（54%）で、GPT-2 生成セットからランダムに選んだ場合のパフォーマンス(66%)はわずかに良かった。

◆ ChatGPT のリリース以来、研究者たちは ChatGPT の学問的課題への対応能力を調査している。
- ➢ Yeadon et al.(2023)は ChatGPT で 300 語のエッセイを 5 本生成した。これは、物理学部の 2 年生を対象とした、物理学のモジュールで出された課題に対する回答である。
  - ✧ 採点者は ChatGPT エッセイに、学部生の平均点に匹敵する 71%の平均点を与えた。ChatGPT エッセイは Grammarly（2%）と Turnitin(剽窃チェックツール)（7%）から盗用の可能性が低いことを示す得点を得た。

◆ 大学入学前の教育環境での ChatGPT の使用に関する研究は少ない。
- ➢ そのような研究の 1 つで、ChatGPT をオランダの全国高校英語試験の多肢選択問題や記述問題に対する回答を生成するために使用した(de Winter, 2023)。
  - ✧ ChatGPT は平均的な生徒と同程度の成績を収め、ある種の問題（例：言語的類推）は他の問題（例：条件推論）よりも得意だった。
- ➢ この発見は、高校生の文章と ChatGPT が作成した文章を区別するのが難しい可能性を提起している。

◆ 高校生の文章と ChatGPT で作成された文章を区別する能力には個人差があることが証明される可能性がある。
- ➢ 潜在的な予測因子の 1 つは、この能力に対する自己報告による自信を指標とした

メタ認知である。(Erickson & Heit, 2015)

- ✧ 先行研究では、自信とパフォーマンスの関係は文脈依存性が高く、同じようなタイプの判断であっても大きく異なることが示唆されている(Fischer & Budescu, 2005)。

- ➢ ChatGPT を使った経験は、さまざまな回答のパターンに気づく助けにもなるかもしれない。ChatGPT の傾向を認識することで、評価者が同じプロンプトに対する複数の ChatGPT の回答を見るテストでの識別精度が上がる可能性がある。

- ➢ 指導経験も重要な可能性がある。例えば、関連する種類の文章を採点する専門知識があれば、他の教師が気づかないような生徒の文章の典型的な特徴に気づくかもしれない。

- ➢ 我々は、高校教師と生徒が、ChatGPT によって書かれたエッセイと高校生によって書かれたエッセイを区別する能力を評価する AI 識別テストに取り組んだ。
  - ✧ ChatGPT の使用経験など、この識別能力を予測する可能性のある要因についても調べた。
  - ✧ 参加者は AI 態度評価テストを受け、教育への影響やどのような使い方が許容されるかなど、アカデミックな場での ChatGPT の使用に関する広範な態度を測定した。

3. Method

3.1. Overview

- ◆ この研究は刺激を作成する刺激準備段階と、教師と生徒が AI 識別テストを受けるテスト段階の 2 つの段階からなる。
  - ➢ テスト段階の目的は、各グループの全体的な正確さを評価し、テストの成功と相関する可能性のある要因を特定することであった。

  - ➢ ChatGPT のリリース前に開発された最近の手法に基づき、参加者は AI が生成したテキストと人間が生成したテキストを読み、その出所を推測した(Gunser et al., 2021; Kobis & Mossink, 2021; Clerwall, 2014)。
    - ✧ 具体的には、2 対のエッセイを並べて提示し、どちらが AI によって生成されたように見えるかを尋ねた。私たちは、高校の教師と生徒をテストし、英語の授業で単位を取るために書いた生徒の文章のサンプルを入手することで、このタスクを高校での文脈に適応させた。

- テスト段階では、AI 態度評価を実施し、AI 識別テストの成績に相関すると思われるいくつかの項目（例：専門分野の知識）について自己申告した。

3.2. Participants

◆ テスト段階では、69 人の高校教師（うち 23 人は英語、45 人は他の教科を教えた）を募集した。ほとんどの教師は女性（54%）、白人（71%）、英語を母国語とする教師（97%）であった。また。高校生 140 人（M age=16.86 歳、女子 44%、白人 57%、英語ネイティブ 96%）を募集した。

3.3. Procedure

3.3.1. Stimulus Preparation Phase

◆ 刺激準備段階は、高校の英語教師と共同でエッセイのプロンプトを作成した。
  - （1）文学エッセイ：なぜ文学は重要なのか？三人称で書いてください。回答は 4〜6 文です。
  - (2)ことわざエッセイ：私たちの世界には多くのことわざがある。例えば、"When life gives you lemons, make lemonade”（災い転じて福となす）や "The early bird catches the worm"（早起きは三文の得）などがある。自身でことわざを選び、その意味を分析してください。三人称で書いてください。回答は 4〜6 文です。

◆ 英語教師は、4 つの英語クラス（2 年生 3 クラス、4 年生 1 クラス）の生徒全員に両方のエッセイを課し、それぞれの課題に対して 97 のエッセイが作成された。生徒には、採点されると告げた。AI の助けなしに書かれることを確実にするため、生徒たちは、個人面談の間に、手書きで書いた。その後、研究アシスタントが手書きの回答をすべて書き写し、以下の基準で回答を除外した。
  - （1）三人称で書く。(2) 指定された長さに収まる。(3) 読みやすい。(4) プロンプトに関連した内容を含める。
    - ✧ これらの除外を適用した結果、文学エッセイ 40 本とことわざエッセイ 42 本が残った（ほとんどのエッセイは三人称を使用していないという理由で除外された。

  - 次に、無作為化ブログラム（http://randomresult.com）を使用して、各セットから 25 のエッセイを選んだ。
    - ✧ ワープロを使って技術的な誤り（単数・複数の一致、スペル、大文字、句読点など）を修正し、明らかな手がかりとならないようにした。

- 2 人の教師が、自分の生徒のエッセイを採点するのと同じ方法で採点した。それぞれの評点を平均して、生徒が書いたエッセイの質を点数化した（mean：85.95 点、SD=6.21 点、range：74.50 to 97.00）。

◆ ChatGPT 刺激は、ChatGPT に同じプロンプトを入力し、プロンプトごとに 25 種類のエッセイができるまで回答を再生成することで準備した。

3.3.2. Testing Phase

◆ テスト段階では、教師と生徒が参加者となった。
- 生徒は、同じ高校の異なる 10 クラスから集めた。この段階には、刺教準備段階に参加した生徒は含まれていない。
- すべての参加者は、まず学校名と教師または生徒としての役割を報告した。そして、ChatGPT の使用経験と、ChatGPT で作成された文章と高校生が作成した文章を区別できるかどうかの自信を示した後、AI 識別テストを行った。

◆ AI 識別テストは 6 組のエッセイから構成された。
- 各組には ChatGPT エッセイと高校生エッセイが 1 つずつ含まれ、各エッセイは適切な 25 のエッセイプールからランダムに選ばれた。プロンプトの順序は、3 つの文学ペアか 3 つのことわざペアのどちらかを最初に表示するように、カウンターバランスをとった。

◆ 参加者はそれぞれのペアのエッセイを読み、ChatGPT によって生成されたと思われるものを選んだ。
- 推測の後、「この答えについてどの程度自信がありますか？」と尋ねられた。（0＝まったく自信がない～100＝非常に自信がある）。
- すべての推測を終えた後、何組のペアが正解したかを 0 から 6 まで推定した。

◆ 次に、参加者は AI 態度評価を受け、ChatGPT を使用してカンニングをする学生がどれくらいいるかを予測し、教育への影響についての懸念と希望を共有し、ChatGPT のさまざまな使用の可能性を評価した（プロンプト: 表 1 参照）。
- この態度評価は、日常的な違反行為（例：学業上のカンニング）に対する人々の評価を測定した過去の手法に基づく(Waltzer & Dahl,, 2021; Waltzer et al., 2023)。

TABLE 1: Summary of AI Attitude Assessment Prompts.

Measure	Prompt	Response options
Estimated cheating	Assuming ChatGPT is freely available, what percent of high school students in the United States would ask ChatGPT to write an essay for them and submit it?	0% to 100%
Concern	How concerned are you about ChatGPT having negative effects on education?	0 = not at all concerned to 100 = extremely concerned
Optimism	How optimistic are you about ChatGPT having positive benefits for education?	0 = not at all optimistic to 100 = extremely optimistic
Comments	(Optional) Do you have any comments? You are welcome to explain your thoughts about any of the above questions.	An open-ended text box
Evaluation of uses	Please rate how good or bad the following actions would be, in your personal opinion. (e.g., a student uses ChatGPT to write an essay for them and submits the directly generated answer)	-10 = really bad to 0 = neutral to +10 = really good

Note: quantitative responses were made using a slider.

◆ 参加者全員は、すでに知っている題材（英文学など）がどの程度あったか評価するよう求められた。教師はまた、英語のクラスを教えた経験について尋ねられた。最後に、参加者は基本的な人口統計学的情報（年齢、性別など）を報告した。

3.4. Data Analysis

◆ 我々は関心のある変数（例：AI 識別テストの全体的な精度）を記述的にまとめた。推論的な検定を使用して、グループ（教師または生徒）、自信、専門知識、ChatGPT への慣れに基づいて AI 識別テストの精度を予測した。また、AI 態度評価への反応をグループ（教師または生徒）に応じて予測した。

- 主要な仮説は、群間比較には Welch の 2 標本 t 検定、その他の正確さの予測因子には F 検定付きの線形回帰モデルで検定した。
- 被験者内の試行ごとの分析には尤度比検定付きの一般化線形混合モデル（GLMMs(Hox et al., 2010)）を使用した。
  - 参加者のランダム切片を使用し、自信とエッセイの質を固定効果として、試行パフォーマンス（正解または不正解）を予測した。

4. Result

4.1. AI Identification Test

4.1.1. Overall Performance

◆ 教師は 70%の確率で ChatGPT によって生成されたエッセイを正しく認識し、学生は 62%の確率で正しく認識した。

- 二項検定の結果、両グループとも偶然の結果を上回った(chance: 50%; teachers: $p < .001$, 95% CI [65%, 74%], students: $p < .001$, 95% CI [58%, 65%])。
  - それでも、完璧な結果にはほど遠く、ほとんどの教師（84%）と生徒（87%）が少なくとも 1 つのトライアルを間違えた。

- Welch の 2 標本 t 検定で、教師は生徒よりも良い結果を出した(t(160) = 2.50, p = .013)。
- 課題前の一般的な自信は、成績を予測しなかった(F(1) = 0.09, p = .760)。
  - 課題終了後でさえ、参加者がいくつ正解したと推定しても、その正確さは予測されなかった(F(1) =0.95, p=.330)。
- どのグループも、英文学の先行経験は成績を有意に予測しなかった(Fs>2.76, ps > .098)。
- 英語教師は非英語教師と同程度の成績だった（t(43) = 0.53, P=.599)。
- ChatGPT を使用した経験の有無も、成績を予測しなかった（Fs< 3.48, ps> .064)。

4.1.2. Trial-by-Trial Performance

- 各試行における参加者の自信は、成績を有意には予測しなかった（D(1) = 0.48, P = .490)。
- 学生のエッセイの質は成績を予測した。
  - 参加者が正解した試行では、参加者が不正解した試行（m=86.72）よりも、刺激準備段階で低品質（m = 85.47）と評価された生徒のエッセイがあった(D(1) = 12.03, p < .001)。

4.2. AI Attitude Assessment

4.2.1. Concerns and Hopes about ChatGPT

- 参加者は、ChatGPT が自由に利用できるのであれば、高校生の約 3 分の 2 がカンニングをするために ChatGPT を利用すると推定した（教師：67%、生徒：64%)。
  - これらの推定値はグループによって有意な差はない(t(147) =1.15, p = .251)。

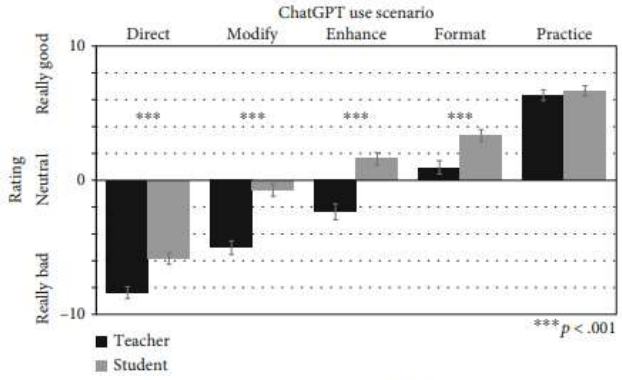
- 0(まったく懸念しない）から 100（非常に懸念する）までのスケールで、教師（m = 70.25）と生徒（m = 51.62）の両方が、ChatGPT が教育に悪影響を与えることについてかなりの懸念を表明したが、教師の方が有意に懸念していた(t(181) = 5.18, p < .001)。
  - 学生は、ChatGPT が教育にプラスになることについて、教師（m = 38.91）よりも楽観的（m=53.13）だった(t(166) = 4.02, p < .001)。

4.2.2. Evaluations of Different Uses of ChatGPT

- 学生にとって ChatGPT を利用することが良いことなのか悪いことなのかについて、参加者の評価は利用方法によって異なっていた（図 3)。
  - 例えば、-10（本当に悪い）から＋10（とても良い）のスケールでは、教師（m = -8.38）と生徒（m = -5.87）は、生徒が ChatGPT を使ってエッセイを書いて提出することはかなり悪いと評価した。

- 一方、生徒がエッセイを書き、エッセイの語彙や構成を質問し、提出することについて、教師はそれを少し悪いと評価し（m = -2.35）、生徒は少し良いと評価した（m = 1.58）。

◆ 教師は、ChatGPT のほぼすべての仮定的な使用について、生徒よりも否定的な評価をした（ts>3.42, ps < .001）。
- 例外は、授業の勉強中に ChatGPT を使って演習問題を生成することだった（両グループとも肯定的）（t(166) = 0.64, p= .525）。



Direct: A student uses Chat GPT to write an essay for them and submits the direct generated answer.
Modify: A student uses Chat GPT to write an essay for their class, then the student edits the output and submits the revised essay.
Enhance: A student writes an essay, asks Chat GPT to improve the essay's vocabulary or structure, and then submits the exact revision.
Format: A student writes an essay, then uses Chat GPT to change it into a specific format (e.g., MLA, APA) and submits the output.
Practice: A student uses Chat GPT to generate practice problems while studying for class.

FIGURE 3: Participant ratings of various ChatGPT uses, grouped by teachers versus students. Bars represent standard error of the mean.

5. Discussion

◆ ChatGPT は、一般に公開されている先行製品をはるかに凌ぐ機能を備えている。一般人も研究者も、個人と社会制度に対するその影響について疑問を呈している。
- このような疑問の多くは、教育における学問的誠実さをめぐるものである（Susnjak, 2022; Roose, 2023）。
- 本研究は、高校におけるこの問題を検討するための第一歩を踏み出すものである。

◆ 我々は、教師と生徒が高校生が書いたエッセイと ChatGPT のエッセイを区別する能力を評価するために AI 識別テストを開発した。
- 平均して、教師（正確さ 70%）は生徒（正確さ 62%）よりわずかに良いスコアだった。
- どちらのグループも偶然（50%）よりも大幅に良いスコアを獲得したが、全体的な成績は、両方のグループが AI 識別テストが非常に難しいと感じたことを示している。

- 本研究の主な焦点は、教師が一般的に生徒のエッセイを評価する責任を負っていることから、AI 識別テストにおける教師の成功に関連する要因を特定することであった。

- 先行研究では、自信がどの程度自分の成績の良い指標となるかは、文脈に大きく依存することがわかっている(Kobis & Mossink, 2021; Wixted &Wells, 2017)。われわれは、現在の状況では、自信は精度の低い指標であり、本質的に役に立たないことを発見した。
  - また、教師たちに ChatGPT の使用経験や英語を教えた経験の有無について尋ねたが、どちらの要素も精度を予測することはできなかった。

- 刺激準備段階の一環として、高校生が書いたエッセイを二人の教師が評価した。より質が高いと判断されたエッセイは、ChatGPT によって生成された文章と区別することが難しく、これは ChatGPT が多くの高校生の文章よりも質が高いとみなされるアウトプットを生成する傾向があることを示唆している。
  - 平均点（83%）を上回ったある教師は、この差異を把握していたようで、”より良く書かれたパラグラフは AI だと思う傾向があった”と述べている。
  - また、曖昧な文章や独特な表現など、生徒のエッセイの指標となるような具体的な側面もあったかもしれない。

- ChatGPT で作成された文章には、参加者の回答に影響を与えるような手がかりがあった。
  - 自由形式のコメントでは、何人かの教師が overall という単語を ChatGPT によって生成されたエッセイであることを示す指標として自発的に扱っていることを指摘した。

- 教師と生徒の ChatGPT に対する意識も調査した。
  - 教師は生徒よりも教育への肯定的な影響に焦点を当てる傾向が低く、楽観的よりも悲観的に多く報告していることがわかった。
    - 対照的に、生徒たちは悲観的と楽観的をほぼ同レベルで回答した。
  - 学生たちはまた、学問的な文脈における ChatGPT のさまざまな利用法を、それほど悪いものではないと評価した。

- この研究にはいくつかの限界がある。
  - この結果は、1 つの教科の狭い範囲のエッセイのみを対象としたものであり、結果の一般化可能性を判断するためには、より多くの研究が必要である。
  - また、ある方法論的決定が結果にどのような影響を与えたかも不明である。

- ✧ 例えば、学生のエッセイの明らかなスペルミスや文法ミスを修正するためにワープロを使用したなど。
- ➢ さらに、ChatGPT のプロンプトに追加情報を与えることがパフォーマンスにどのような影響を与えるか（例えば、高校生のように書くように要求したり、特定の経過的な語を避けるように要求したり）を調べることも重要である。

◆ 生徒の文章と ChatGPT のようなチャットボットが作成した文章を区別することは、すでに重要な課題となっており、AI が進歩し続けるにつれてさらに困難になる可能性がある。

- ➢ この問題を克服するためには、コンテンツ分野の専門知識や ChatGPT に精通しているだけでは不十分である。
- ➢ ChatGPT の全面禁止は効果がなく、ChatGPT のリスクを最小限に抑えながら、その利点を理解し活用する方法を見つけることが、より実りあるアプローチになるのではないか。

◆ 教育的文脈における ChatGPT の役割について人々の考えを理解することは、生徒の学習に対する意味合いだけでなく、公平性を含むより広範な道徳的問題にとっても重要である。

- ➢ 実証的なエビデンスと活発な議論は、この新しい教育の展望を切り開く上で重要な役割を果たすだろう。

6. Conclusion

◆ 本調査では、ChatGPT が作成したエッセイと高校生が作成したエッセイを見分ける能力は、高校生よりも高校教師の方が若干優れていることがわかった。

- ➢ しかし、教師でさえ、特に生徒のエッセイがよく書けている場合、このタスクにかなり苦労していた。
- ➢ ChatGPT で作成されたエッセイを見分ける能力に自信のある教師は、自信のない教師よりも識別に優れていることはなく、ChatGPT の経験や関連した専門知識は、この能力とも無関係だった。
- ➢ 私たちの態度評価によると、高校生は ChatGPT で書かれたエッセイを提出することに一般的に否定的な見方をしていたが、このような行為や他の潜在的な倫理違反については、教師よりも否定的な見方をしていなかった。
- ➢ 生徒たちは、ChatGPT が教育に有益であると楽観的に答える傾向が教師よりも強かった。

◆ これらの結果を総合すると、ChatGPT は高校における学術的誠実性を脅かすものであ

り、新たな政策や社会規範の策定が必要であることが示唆される。

- 我々の発見は、教育システムがこの強力で急速に進歩するテクノロジーにどのように適応していくべきかについて、様々な重要な問題を提起している。