

自己調整学習における自動足場はずしのための計画と 振り返りの自動評価手法の開発

廣瀬 伸行^{1,2,a)} 白松 俊^{2,b)} 奥原 俊^{3,c)}

受付日 2022年8月15日, 採録日 2023年1月19日

概要: 自己調整学習を支援するためには, 学習者自身が学習状況を把握しやすくするために, 振り返りや計画を支援する機能が重要となる. 学習者数が多い場合に個別の状況に対応するには, 足場かけや足場はずしを自動設定できることが望ましい. さらに, アドバイスも自動化できると教員の負担を減らすことができる. そこで, 学習者の振り返りと計画の記述を自動評価し, 足場はずしのレベルを自動設定する手法を開発した. 具体的には, OpenAI の大規模言語モデル GPT-3 を再学習するために, 実際の授業から得られた学習者の振り返りと計画の記述, および教員が手動評価したデータを用いた. 提案手法を用いて, 実際の授業における学習者の振り返りと計画を自動評価し, その性能を検証した. その結果, 教員の手動評価と比較して, 記述の具体性については順位相関係数 +0.815, 躓き度については順位相関係数 +0.876 であった. 十分に実用に耐える性能と考えられるが, 躓き度については経緯を考慮した推定ができない場合があるなどの課題も明らかになった.

キーワード: 学習支援システム, 自己調整学習, 自動評価, GPT-3

Developing Method for Automatic Evaluation of Planning and Reflection for Automatic Scaffolding Removal in Self-Regulated Learning

NOBUYUKI HIROSE^{1,2,a)} SHUN SHIRAMATSU^{2,b)} SHUN OKUHARA^{3,c)}

Received: August 15, 2022, Accepted: January 19, 2023

Abstract: In order to support self-regulated learning, it is important to have functions that support reflection and planning in order to facilitate the learner's own understanding of their learning status. In order to respond to individual situations when the number of learners is large, it is desirable to be able to set scaffolding and scaffolding removal automatically. In addition, the automatic advice function reduces the workload of the teacher. Therefore, we developed a method that automatically evaluates learners' reflections and descriptions of their plans and automatically sets the level of scaffolding removal. Specifically, learner reflections and planning statements from actual classes and data from teachers' manual evaluations were used to retrain OpenAI's large-scale language model GPT-3. We tested the performance of the proposed method by automatically evaluating learners' reflections and plans in an actual class. The results showed a rank correlation coefficient of +0.815 for the specificity of the description and a rank correlation coefficient of +0.876 for the degree of stumbling, compared to the teachers' manual evaluation. The performance of the system is considered sufficient for practical use, but there are some issues that need to be addressed, such as the inability to estimate the degree of stumbling that takes into account the history of the stumbling.

Keywords: learning support systems, self-regulated learning, automatic evaluation, GPT-3

¹ 愛知産業大学
Aichi Sangyo University, Okazaki, Aichi 444-0005, Japan
² 名古屋工業大学
Nagoya Institute of Technology, Nagoya, Aichi 466-8555, Japan
³ 三重大学
Mie University, Mie 514-8507, Japan
^{a)} hirose@asu.ac.jp
^{b)} siramatsu@nitech.ac.jp
^{c)} okuhara@eng.mie-u.ac.jp

1. はじめに

学習者の学習支援を自動化する場合, 学習者個別の学習状況を自動的に定量化する手法が課題となる. 学習者の主体的な学習には, 学習者が自身の学習状況を把握するメタ認知が重要である [1]. そこで, 学習者に振り返りと計画

を促す足場かけ、足場はずしなどの学習支援が行われる [2]。足場かけ、足場はずしの設定やアドバイスの選択を自動化をする場合、学習者の振り返りと計画の記述内容を自動的に評価して定量化する必要がある。

そこで、本研究は以下の2つを目的とする。

- (1) GPT-3 を用いて、具体性と躰き度について、教員の手動評価を模倣した自動評価ができる。
- (2) GPT-3 で自動評価した値を用いて足場はずしを自動的に設定できる。

以下に、前提となる学習支援における評価と足場かけ、足場はずしの関係と必要性を説明する。そして、自動評価を実装する開発中の学習支援システムと、本研究の目的である自動評価の必要性について説明を行う。

1.1 具体性と躰き度を評価する必要性

振り返りと計画の具体性、および、躰き度の評価をしなければならない理由は、学生の主体的な学習のために、どの程度の支援が必要なのかを判断するためである。

オンライン教育では、学習者が自宅などから参加して学習活動を行うほかに、参加しない時間における主体的な学習活動も重要となる。主体的な学習者は、自身の学習状況を常に把握して、自身の学習方略を自己調整しているとされる [3], [4]。つまり、オンライン教育などある程度主体的な学習が求められる学習活動では、学習者が自身の学習状況を把握して計画的に見通しを立て、学習活動を調整していくことが重要である [5]。

振り返りと計画が具体的に記述できていない学習者は、自身の学習状況の把握や学習の見通しがうまくできていない。さらには自身の躰きにも気が付いていないことがある。したがって、振り返りと計画が具体的に記述できるよう支援を行う必要がある [6], [7]。

1.2 足場かけと足場はずしの必要性

振り返りや計画の記述は、学習者が自身の状況を吟味を促し、学習の工夫を促す [5]。足場はずしは、主体的な吟味や工夫を修得する機会として必要である。しかし、うまく記述ができない学習者には、足場かけによる記述支援が必要である。さらに、教員による手動評価や機械学習による自動評価のためには、できる限り学習状況が記述された文章であることが望ましい。そのために、足場かけ、足場はずしを行う必要がある。

足場かけ 足場かけは、一定の学習状況を得るための記述支援措置とした。そのために、振り返りや計画について、ヒントや選択肢、定型文を用いて、ある程度の型にあてはめられた記述をさせる。つまり、何をどのように記述すべきか分からない学習者からも学習状況を読み取りやすいテキストデータが得られる。しかし、型に当てはめようとするため学習者個別の状況に応じた主体的な工夫を阻害する恐れ

もある。そのために記述内容の状況に応じた足場はずしを必要とする。

足場はずし 足場はずしは、学習者個別の吟味と工夫を含んだ学習状況を得るための措置とした。そのために、自由記述欄を用いることで、学習者が自身の状況を吟味した記述させる。つまり、学習者個別の学習状況に応じて吟味と工夫を含んだテキストデータが得られる。しかし、型にあてはまらないテキストデータのため学習状況を読み取ることが比較的難しい。また、何をどのように記述すべきか分からない学習者については、教員が意図しないデータを含む可能性が高まる。そのために、記述内容の状況に応じた足場かけを必要とする。

なお、本研究では、足場かけ、足場はずしの妥当性や、主体的な記述ができるようになったかどうかの評価についてはできていない。今後の研究課題である。

これらの足場かけ、足場はずしは学習者個別の学習状況に応じて判断することが重要となる。

1.3 学習支援システムの概要

図1は、筆者らが開発している学習支援である。図1の上部が提案手法実装前、下部が実装後を示す。筆者らは、以前開発した学習支援 [7] を改善するため、学習者の振り返りと計画の記述の具体化、および記述に伴う学習状況把握から自己調整を支援するシステムを開発している。

具体的には、学習者の振り返りと計画の支援として、3つの足場はずしレベル (Lv.1-3) を設定した。

- (1) **Lv.1** 支援が手厚。ボットが記述内容をサポート。
- (2) **Lv.2** 支援が手薄。記述内容に関するヒント提示。
- (3) **Lv.3** 支援が無い。学習者が自力で記述。

図2は、図1に示した専用アプリの画面である。図2の専用アプリは提案手法の実装前後で共通であり、本研究の実験に使用された。図2の「TASK」が学習支援の実装

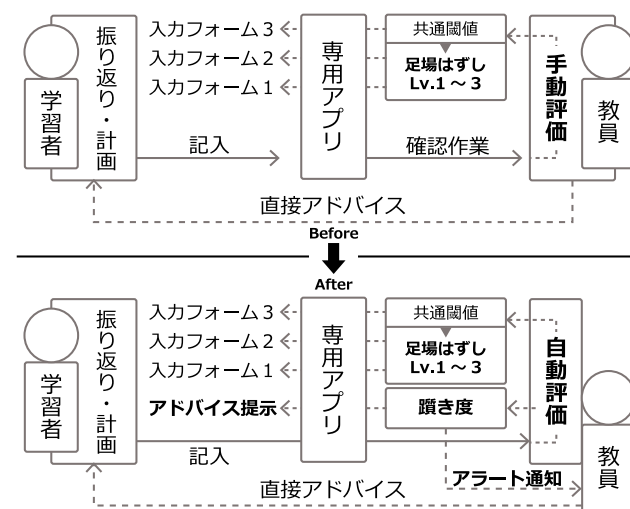


図1 学習支援の概要

Fig. 1 Outline of learning support.

部分である。専用アプリは、授業課題の制作環境と学習支援環境を統合した Web アプリである。「TASK」以外の「IMAGE」「CASE」「TEST」は、授業課題であるデジタルサイネージを制作するための機能やツールを備える。「TASK」には、学習支援として、振り返りや計画の記入、記入有無の表示、記述内容の再表示、アンケート回答、課題タスクを管理する機能を備える。

学習者は、振り返りや計画の記述を図2に示した「TASK」内の赤枠点線内に提示される入力フォームを行う。入力フォームは3種類あり、授業回ごとに学習者個別に設定された足場はずしレベル (Lv.1-3) に応じて選択され提示される。足場はずしレベルごとに提示される入力フォーム (1-3) の例は4章実験方法で示す。

振り返りや計画にされた記述内容は専用アプリ内に保存される。教員は、次の授業回の足場はずしレベルを設定するため専用アプリ内に保存された記述内容について目視と手作業で具体性の評価を行う。そして、教員は必要に応じて学習者にアドバイスをを行う。

1.4 自動評価の必要性

図1の上部に示す手動評価による学習支援の運用について、大学などの毎週実施される授業では、学習者が多く、さらに複数の授業を担当する教員にとって多大な負担となった。そのために、学習支援のためのアドバイスについても緊急性が高い場合など最小限となり、提示するタイミングが遅れるなどの問題が生じた。そこで、学生個別の学習支援を充実しつつ教員の負担を増やさないために、足場はずしレベルの自動設定や提示するアドバイスの選択を自動化する方法が必要となった。



図2 開発中の学習支援画面

Fig. 2 Learning support screen under development.

図1の下部は、自動化したときの学習支援を示す。学習者の振り返りと計画の記述内容から具体性と躰き度を自動評価する。自動評価によって以下2つの自動化を実現する。

- (1) 具体性の自動評価 足場はずしレベルの自動設定を行うために、学習者が記述した振り返りと計画から具体性を自動評価を行う必要がある。そのために、振り返りを細分化した4項目と計画を細分化した4項目について具体性を自動評価する。その合計値によって足場はずしレベルを設定する。
- (2) 躰き度の自動評価 今後の課題であるアドバイス自動選択と教員へのアラート自動通知を実現するために、躰き度を自動評価する必要がある。そのために、振り返り4項目と計画4項目の全体の記述内容について躰き度を自動評価する。

以上2つの自動化を実現することで、さらに入力フォームの提示と同時に学習者個別にアドバイスを自動提示することも実現できる。

図3は、図1の下部に示した自動評価の詳細を示す。学習者が記述する振り返りと計画の記述内容は、振り返り4項目、計画4項目の計8項目で構成される。図3点線部に示した項目の自動評価を実現するため、学習の振り返りと計画の8項目の記述内容から具体性と躰き度を自動的に評価する手法が課題となる。

そこで、本研究では、OpenAIの大規模言語モデル GPT-3 [8] の再学習を行うことで自動的な評価を実現する。自動的な評価を実現するための自動評価用モデルは、GPT-3の事前学習モデル curie をトレーニングデータで再学習する。再学習した自動評価用モデルを用いることで、学習者の振り返りと計画の記述内容から具体性と躰き度を自動評価する手法を提案する。

本研究は、自動評価用モデルによる自動評価の有効性について、自動評価用モデルによる自動評価と教員による手動評価を比較した検証を行う。

先行研究では、先行研究における限界を示し、教員の評

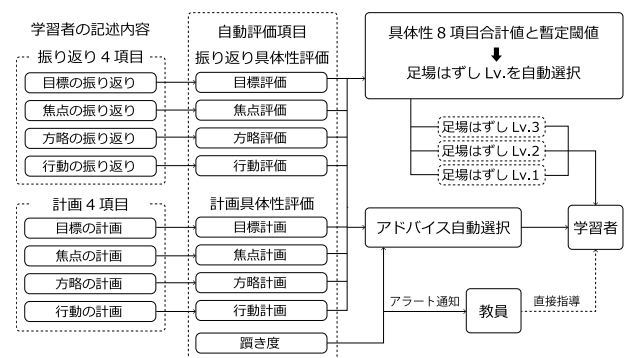


図3 学習支援の自動評価したい部分

Fig. 3 The part of the learning support we want to evaluate automatically.

価を模倣した自動評価手法に GPT-3 を用いる新規性を示す。提案手法では、自動評価モデルを得るための GPT-3 の再学習と、トレーニングデータの作成方法を示す。実験方法では、実際の授業からトレーニングデータ、および、テストデータを得た方法を示す。評価方法では、交差検証法を用いて自動評価した実験データを検証する方法を示す。実験結果では、具体性の自動評価に順位相関係数 + 0.815、躓き度の自動評価に順位相関係数 + 0.876、足場はずしレベルの自動設定に順位相関係数 + 0.804 が得られたため、それらの詳細と、評価が一致した事例と誤差が生じた事例を示す。考察では、実験結果から本研究の本研究の目的 1 と 2 をある程度達成できており、実用に耐える性能と結論した考察と、考察から得られた今後の課題を示す。まとめでは、考察結果を総括し、結論と今後の課題を示す。

2. 先行研究

学習の振り返りの記述内容を評価して定量化する手法として、ファイルサイズに着目した事例がある [9]。ファイルサイズが定量化の指標となりうる可能性について、学習者が手書した振り返りの記述内容を機械的に読み取ったときのファイルサイズと学習成果評価との関係と傾向から明らかにしている。しかし、文字数やファイルサイズなどの指標を用いる従来手法は、事例の中でも説明されているように、学習成果との関係性が示されたのみで因果関係が明らかでなく、おそらくは本研究が必要とする記述内容の具体性や躓きを評価するのに適した手法ではない。

学習者のノート記録内の単語数と教員が提示した単語や単語数との比較結果を特徴量として、教員によるノート評定との関係からテストの得点を予測する可能性を示した事例がある [10]。本研究が焦点をあてる記述内容の具体性や躓き度を評価するものではないものの、学習者のノート記録に出現する単語や単語数に着目し、教員の付けたノート評定の関係から推定したテストの得点として定量化されている。具体性の躓き度の評価にも応用できる可能性があり事例として参考にできる。しかし、本研究の自動化に応用するには、学生の記述内容と教員が提示する内容を解析して比較を特徴量として評価するための事前事後のデータ準備を授業回ごとに行う手間が負担となる。

機械学習モデルを作成して学習者の振り返り内容を分類する手法を示した事例がある [11]。自然言語処理の手法を用いたテキストデータの分類モデルを再学習し、学習者の振り返りの記述内容から学習者の振り返り状況を分類することで、学習者の自己評価と教員の評価を支援する手法を示している。学習者の振り返り記述内容と分類を再学習したモデルを用いる評価手法が参考にできる。本研究では振り返りと計画の記述内容と教員が評価した具体性と躓き度の数値をトレーニングデータに用いて再学習したモデルに

よる自動的な評価手法を提案する。

3. 提案手法

学習者の振り返りと計画の記述内容から具体性と躓き度を自動評価する。以下 2 つの目的を実現するために、大規模言語モデルの再学習を用いる手法を提案する。

- (1) 具体性と躓き度について、教員の手動評価を模倣した自動評価ができる。
- (2) 自動評価した値を用いて足場はずしを自動的に設定できる。

以下の 3 点は、具体性や躓き度の評価に GPT-3 を用いる理由である。

- (1) ドメイン固有ではない予測困難な語彙にも対応できる。GPT-3 は大規模なコーパスで事前学習されているため、学生が用いる予測不能な語彙にも対応可能である。
- (2) 再学習 (fine-tuning) が容易である。たとえば BERT の場合、分類問題か回帰問題かによって再学習の方法が変わってくるが、GPT-3 であればどんな問題であっても出力をテキスト形式で扱うため、訓練データの設計が容易である。
- (3) GPT-3 の意味解釈の性能の良さ。具体性や躓き度の評価は、振り返りや計画に記述されているテキストから、学習に関する重要な点と、それらの関係性を捉えて解釈する必要がある。GPT-3 は、そのような意味解釈を高精度に行える。

3.1 自動評価用モデルの作成手順

図 4 は、自動評価用モデル作成の手順を示す。トレーニングデータは、学習者が記述した学習活動の振り返りと計画内容に対して授業担当教員が手作業で具体性と躓き度を評価した結果を用いる。

具体的には、トレーニングデータは、図 3 の学習者の記述内容に示した振り返り 4 項目 (目標の振り返り、焦点の振り返り、方略の振り返り、行動の振り返り) と計画 4 項目 (目標計画、焦点計画、方略計画、行動計画) の計 8 項目について、図 3 の自動評価で示す 9 項目 (目標評価具体

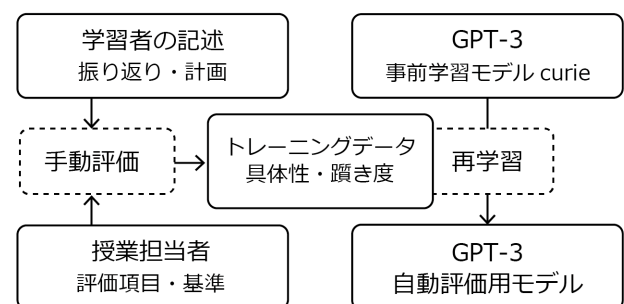


図 4 自動評価用モデルの作成

Fig. 4 Creating model for automatic evaluation.

性、焦点評価具体性、方略評価具体性、行動評価具体性、目標計画具体性、焦点計画具体性、方略計画具体性、行動計画具体性、躓き度)について教員が手作業で評価した結果を用いる。

3.2 教員の手動評価によるトレーニングデータの作成

具体性8項目の評価は、記述の具体的なものを「5」とし、具体的でないものを「1」の5段階評価とした。躓き度の評価値は、躓きが大きいものを「5」とし、小さいものを「1」の5段階評価とした。具体的な評価基準は、付録のA.1に示した。なお、評価基準どおりに教員が評価できているかについて、評価の妥当性の検証はできていないため、今後の課題とする。

表1に、振り返りの記述に対して教員が手動評価した一部例を示す。焦点の振り返り(焦点評価)に対する具体性評価5から1の例を示した。

表1の各評価について、ループリックに照らした評価を示す。

評価5 「設定した条件が出てこない」という結果を回避するために、データの作り方に関係があることに注意した記述が評価された。

評価4 良い結果を得るために「条件を可能な限り散らした」「画像に合った条件を考えた」が寄与していると注意した記述が評価された。

評価3 「様々なパターンを考え」ることに注意した記述が評価された。

評価2 「ユーザー目線」とすることに注意した記述が評価された。

評価1 記述内容が注意点として評価されなかった。躓き度1とした表2の事例は、振り返りと計画が具体的になっていない記述が評価された。

表2に、トレーニングデータの例を示す。「prompt」に振り返りと計画の記述、「completion」に教員の付けた評

価を記述した。

表2の躓き度について、ループリックに照らした評価を示す。

躓き度1 振り返りと計画が具体的になっていない記述が評価された。

3.3 事前学習モデルの再学習

以上のトレーニングデータを用いて GPT-3 の事前学習モデルを再学習することで自動評価用モデルを得る。再学習は、GPT-3 の事前学習モデル curie を用い、バッチサイズは2、エポック数は6とした。

4. 実験方法

再学習に用いるトレーニングデータ、および、評価に用いるテストデータを実際の授業から得るために実験を行った。以下にデータを得るために実施した実験方法について示す。

4.1 実験から評価までの手順

- (1) トレーニングデータの作成. 実験対象の実際の授業から学習活動の振り返りと計画を取得。記述内容について教員が手作業による具体性と躓き度の評価を行う。
- (2) 自動評価用モデルの作成. トレーニングデータを用いて自動評価用モデルを作成。
- (3) 交差検証. 自動評価用モデルを適用した自動評価の結果と教員による手動評価を検証。

表2 トレーニングデータの例
Table 2 Training data sample.

{
"prompt":
振り返り_目標: 「あまり進められなかった」,
振り返り_焦点: 「深く考えすぎない」,
振り返り_方略: 「うまくいきつつある」,
振り返り_行動計画: 「空き時間に取り組んだ」,
計画_目標: 「やれる時に一気にやる」,
計画_焦点: 「効率」,
計画_方略: 「考えすぎずにテキパキとこなす」,
計画_行動計画: 「授業中や空き時間」
-> ",
"completion":
目標評価具体性=1, 焦点評価具体性=1,
方法評価具体性=1, 行動評価具体性=1,
目標計画具体性=1, 焦点計画具体性=1,
方法計画具体性=1, 行動計画具体性=2,
躓き度=1
END "
}

表1 教員の手動による具体性評価の例(焦点振り返り)

Table 1 Example of a teacher's manual concreteness evaluation (Focused reflection).

焦点振り返り記述	評価
一番注目した所は、デシジョンツリーを作成する時、何回か設定した条件が出てこない現象が起こったので、どのように作成したら、設定した条件が出てくるか考えながら作成した部分が一番注目した。	5
効果があったと思う方法や考え方は、条件を可能な限り散らした。興味を惹くようなところは特になかった。課題をやってみた印象は、画像に合った条件を考えたこと。	4
様々なパターンを考えてデシジョンツリーの行を増やすことに注目した。	3
ユーザー目線の学習用データの入力	2
ディジョンツリー	1

4.2 実験対象

対象授業 デジタルデザイン表現

被験者 大学生 156 名

演習課題 デシジョンツリーで動くデジタルサイネージの作成

対象授業回 10 回目, 12 回目

振り返りと計画の提出数 253 件 (10 回目 128 件, 12 回目 125 件) 未提出は含まない

毎回の授業において 156 名の受講者に振り返りと計画の記入を求めた。本研究では, 10 回目と 12 回目の 253 件の学習活動の振り返りと計画の内容を実験対象および実験用のトレーニングデータとして使用した。

10, 12 回目を対象にした理由を述べる。10 回目は, 足場はずしレベル (Lv.) ごとの入力フォームの調整を終え, 全員を足場はずし Lv.3 に統一したタイミングであった。そのため教員の手作業による評価が安定すると考えた。11 回目は, 担当教員である筆者が手作業で評価を行う前に, 筆者が練習で試した自動評価の結果に触れてしまっていたため, あらかじめ自動評価の結果を知っている筆者が手動評価を行うと, 自動評価の結果を意識せずに本来の手動評価を行うことが困難であるため使用しなかった。13 回目は, 11 回目と同じ理由で使用しなかった。

以下に, 12 回目の記述に使用された足場はずし Lv. の件数を示す。足場はずし Lv.3 : 16 件, Lv.2 : 63 件, Lv.1 : 46 件。

4.3 振り返りと計画のサイクルと足場はずしレベルの設定

図 5 に授業回ごとに実施した振り返りと計画のサイクルを示す。たとえば 10 回目の振り返りは, 9 回目に建てた計画に基づいて記述する。そして, 10 回目の計画は, 振り返りと 10 回目の授業内容に基づいて記述する。なお, 振り返りと計画のサイクルにおいて提示される入力フォームの種類については, たとえば, 10 回目に記述するときは, 9 回目の記述内容の具体性評価に基づいた足場はずしレベルによって定まる。

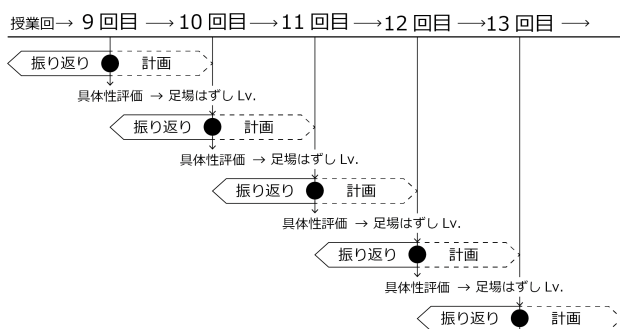


図 5 授業回ごとの振り返りと計画のサイクル

Fig. 5 Reflection and planning cycle for each class session.

以下に 10, 12 回目に関連する可能性のある 9 回目から 13 回目までの授業内容を簡潔に示す。振り返りと計画の記述内容は授業進行に沿って記述されることが想定される。たとえば, 10, 12 回目の振り返りは 9, 11 回目に記述した学習計画に対する自己評価であり, 10, 12 回目の計画は 11, 13 回目までの学習計画となる。

9 回目 デシジョンツリーの調整

10 回目 キャッチコピーの考案

11 回目 キャッチコピー表示とレイアウト

12 回目 デジタルサイネージの動作設定

13 回目 画像表示と切り替えアニメーション

足場はずしレベルの設定は, 図 3 の学習者の記述内容に示された振り返り 4 項目と計画 4 項目について各項目の具体性評価値の合計と暫定閾値 (Lv.1 : 合計値 < 15, Lv.2 : 合計値 < 25, Lv.3 : 合計値 ≥ 26) によって設定される。なお, 暫定閾値の適性検証は今後の研究課題である。

4.4 入力フォームと足場はずしレベル

図 6 に足場はずしレベル (Lv.) ごとの入力フォームの一部を示す。図 6 は, 上から順に足場はずし Lv.3, 2, 1 と並ぶ。すべての入力フォームは, 図 3 に示した振り返り 4 項目 (目標の振り返り, 焦点の振り返り, 方略の振り返り, 行動の振り返り) と計画 4 つの項目 (目標計画, 焦点計画, 方略計画, 行動計画) について学習者が記述する。

以下に各項目の説明を示す。

1. 目標の振り返り (目標評価) 前回計画した目標の妥当性と, どこまで到達したか, 何ができなかったのかについて自己評価する。
2. 焦点の振り返り (焦点評価) 前回計画した焦点の妥当性と, 取り組みから得られた結果と焦点を関係付けて自己評価する。
3. 方略の振り返り (方略評価) 前回計画した方法の妥当性と, 得られた結果と方法を関係付けて自己評価する。
4. 行動の振り返り (行動評価) 前回計画した行動の妥当性と, 得られた結果と行動を関係付けて自己評価する。
5. 目標計画 次回までの到達目標について, 前回の自己評価と現状を踏まえて段階的に計画する。
6. 焦点計画 課題から着目したポイントを明確にし, 自身の興味関心などと関連付けて計画する。
7. 方略計画 課題の方法や考え方, 試みについて, 課題内容や焦点, 自身の得意不得意などと関連付けて計画する。
8. 行動計画 課題をいつ実施するのか, 次回までの自身の予定と関連付けて計画する。

以下に, 図 6 の入力フォームの違いを示す。

足場はずし Lv.3 の入力フォーム 学習者がほぼ自力で記

振り返りをします。前回以降の課題取り組みについて、以下の項目を自己評価をしましょう。

1. 前回の計画に対してどこまでできましたか

2. 課題に取り組んで一番注目したことは

3. 計画した方法や考え方はうまくいきましたか

4. いつ、どんな時に課題に取り組みましたか

次へ 足場はずし Lv.3 入力フォームの一部

今回の課題について次回（授業前）までの計画をしましょう。

5. 今回の課題はどんな段階に分けてやると思いますか？
例えば、まずは画像の表示設定や切り替え効果を試してから、Excelを30行まで進めるとかな。
どうかな。自分なりのステップを作ってみよう。

次へ 足場はずし Lv.2 入力フォームの一部

オートTA: 次は2つ目の記入項目ね。
課題をやってみたと思うんだけど、
そのとき、やってみた方法や考え方で、課題を進めるのに役立ったことや効果があったと思うことは何かある？

あった なかった

ユーザー: なかった

オートTA: そうか。問題ない。そういうときもある。

オートTA: やってみた方法や考え方で、もっと試したいことや、もっと知っておきたいこと、やらなさやと思うことは何かあった？

あった なかった

ユーザー: あった

オートTA: いいね。どんなことをもっと試したり、知りたいと思ったの？

送信

オートTA: 2つ目の記入項目、前回の課題で一番注目したことについて追加記入して送信してね。
送信したら、次の記入項目だよ。

確認送信 足場はずし Lv.1 入力フォームの一部

図6 足場はずしレベルごとの入力フォーム（上からLv.3, 2, 1）
Fig. 6 Forms per scaffold removal Lv (Lv.3, 2, 1 from the top).

述する。振り返り4項目と計画4つの項目の問いと入力欄が提示される。

足場はずし Lv.2 の入力フォーム 記述内容を示唆する支援がある。Lv.3の入力欄に加え、振り返りには前回の記述内容、計画には授業内容に沿ったヒントが提示される。

足場はずし Lv.1 の入力フォーム 記述内容の定形文を示

す支援がある。ボットとの対話で、学習者が選択肢を選び、選択肢に応じた定形文があらかじめ示された入力フォームが提示される。

10回目は、全員が足場はずしLv.3のフォームで記入され、12回目は、11回目の記述内容によって学習者個別に足場はずしレベルが設定され、学習者個別に足場はずしレベルに応じた種類の入力フォームで記述された。

5. 評価方法

本研究の目的1：教員が手作業で評価した手動評価を模倣することができているか、および、目的2：自動評価した値を用いて足場はずしを自動的に設定できるかどうかについて、実験で得られたデータを用いた交差検証法によって評価する。

以下に示す(1)から(3)の3つの自動評価は、本研究の目的2つに対応する。3つの自動評価について、交差検証法を用いた検証を行う。これにより、本研究の2つの目的について評価する。なお、提案手法の評価軸は、自動評価と教員による手動評価の相関係数であり、交差検証法(Cross-validation)で性能評価するのが一般的である。

(1) **具体性の自動評価**。目的1：学習者が記述した振り返りと計画の記述内容について具体性を自動評価したときの性能を評価する。振り返り4項目と計画4項目の計8項目を対象に項目ごとの記述内容を自動評価する。

(2) **躰き度の自動評価**。目的1：振り返り4項目と計画4項目の全8項目の記述内容を総合して躰き度を自動評価したときの性能を評価する。なお、本研究では躰き度のみを評価する。なぜなら、躰き度によるアドバイス自動選択と教員へのアラート自動通知については開発中であり実装されていないためである。

(3) **足場はずしレベルの自動設定**。目的2：自動評価した具体性8項目の合計値を暫定閾値（手動評価と自動評価共通）に代入して足場はずしレベルを設定する。足場はずしレベルの自動設定について性能評価を行う。

表3は、実験用トレーニングデータに含まれた手動評価による評価項目と評価値の度数分布を示す。本研究では10, 12回目の手動評価のみを使用したため、各項目とも評価値5の件数が少ない。

表4に交差検証用のグループ分けを示す。グループ分けは、手動評価の足場はずしレベルの分布が大きく偏らないようにした。データの先頭から順にABCDEの各グループにLv.別に割り振った。

交差検証を行うため、各グループのデータを除外した5つの実験用トレーニングデータを作成し、そこから5つの交差検証用の自動評価用モデルを再学習した。具体的には、Aグループ51件を除いた自動評価用モデル「Excluding A」、Bグループ51件を除いた自動評価用モデル「Ex-

表 3 手動評価された評価項目と評価値の度数分布

Table 3 Frequency of each evaluation item/value in the training data.

評価項目	1	2	3	4	5
目標評価具体性	69	57	82	41	4
焦点評価具体性	74	61	73	39	6
方略評価具体性	143	59	32	18	1
行動評価具体性	42	124	69	15	3
目標計画具体性	60	58	75	52	8
焦点計画具体性	58	68	79	43	5
方略計画具体性	55	63	97	34	4
行動計画具体性	35	126	67	3	2
躓き度	87	49	46	37	34

表 4 交差検証用グループの設定

Table 4 Setting up groups for cross-validation.

Groups	Total number	Lv.1	Lv.2	Lv.3
A	51	15	27	9
B	51	15	27	9
C	51	15	26	10
D	50	14	27	9
E	50	14	27	9
Total	253	73	135	45

cluding B], C グループ 51 件を除いた自動評価用モデル「Excluding C」, D グループ 50 件を除いた自動評価用モデル「Excluding D」, E グループ 50 件を除いた自動評価用モデル「Excluding E」を再学習した。再学習時のパラメータは、事前学習モデル curie を用い、再学習時のバッチサイズは 2, エポック数は 6 とした。また、再学習によって得られた自動評価用の GPT-3 モデルを適用して自動評価する際、ランダム性を表すパラメータ temperature は 0 に設定した。

交差検証用の自動評価結果を得るため、交差検証用の自動評価用モデルで対応するグループの自動評価を行った。具体的には、「Excluding A」モデルで A グループ 51 件、「Excluding B」モデルで B グループ 51 件、「Excluding C」モデルで C グループ 51 件、「Excluding D」モデルで D グループ 50 件、「Excluding E」モデルで E グループ 50 件を自動評価した。

足場はずしレベルの自動設定は、自動評価された具体性の合計値と暫定閾値（手動評価と同じ閾値）によって設定された。

6. 実験結果

本研究の 2 つの目的ために以下の 3 つを交差検証法によって評価した結果を示す。自動評価の性能を示すために、自動評価と手動評価について、ノンパラメトリック（スピアマン）の順位相関分析を行った。

(1) 具体性の自動評価（目的 1）：順位相関係数 +0.815

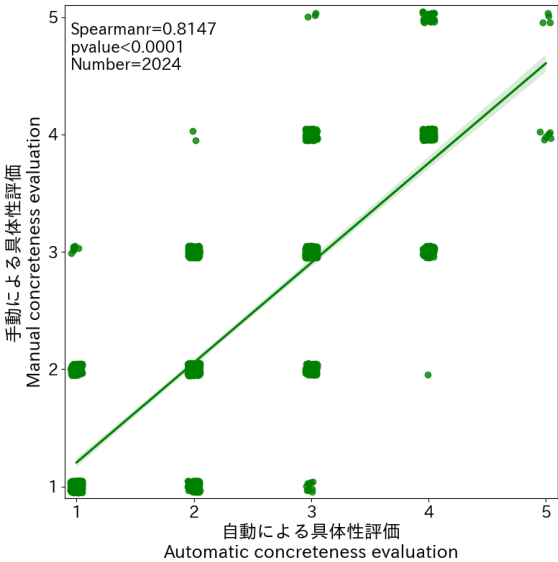


図 7 すべての具体性評価に関する手動評価と自動評価
Fig. 7 All concreteness evaluations by manual and automatic.

- (2) 躓き度の自動評価（目的 1）：順位相関係数 +0.876
- (3) 足場はずしレベルの自動設定（目的 2）：順位相関係数 +0.804

以下に、自動評価の値と手動評価の値の散布図を示す。ただし、評価値が離散的な整数値であるため、そのまま散布図を描くと単なる格子点となり、標本の密度を読み取ることができない。そこで、自動評価値、手動評価値ともに $[-0.05, +0.05]$ の一様乱数を加算して表示している。

なお、以下に示す実験結果の順序は、具体性の自動評価、足場はずし Lv. の自動評価、躓き度の自動評価とした。足場はずし Lv. は、具体性の自動評価の値によって設定されるためである。

6.1 具体性の自動評価

図 7 からは、振り返りと計画の計 8 項目全体の具体性評価について手動評価と自動評価の順位相関係数 +0.815 ($p < 0.0001$, $n = 2024$) が示された。

6.1.1 振り返り 4 項目と計画 4 項目の項目別の結果

図 8 は、振り返り 4 項目と計画 4 項目ごとの具体性評価について手動評価と自動評価の関係を示す。

目標評価具体性（図 8, a）の順位相関係数 +0.855 ($p < 0.0001$, $n = 253$) が最も相関が強く、焦点評価具体性（図 8, b）の順位相関係数 +0.854 ($p < 0.0001$, $n = 253$)、方略評価具体性（図 8, c）の順位相関係数 +0.777 ($p < 0.0001$, $n = 253$)、目標計画具体性（図 8, e）の順位相関係数 +0.841 ($p < 0.0001$, $n = 253$) が正の相関を示した。

一方、行動評価具体性（図 8, d）の順位相関係数 +0.647 ($p < 0.0001$, $n = 253$) は 8 項目の中で最も相関が弱く、行動計画の 4 項目内では、行動計画具体性（図 8, h）の順位相関係数 +0.708 ($p < 0.0001$, $n = 253$) が最も相関

が弱い。

表5は、具体性評価項目ごとの誤差（自動評価値－手動評価値）の件数を示す。すべての評価項目について誤差±1は一定数あったことが読み取れる。さらに、誤差±2に注目したとき、振り返りでは、目標評価具体性と焦点評価具体性について、自動評価が手動評価よりも高く評価する場合があった。方略評価具体性と行動評価具体性誤差では、手動評価よりも低く評価した事例が3件、高く評価した事例が2件あった。計画に関する4項目では、手動評価よりも具体性を2高く、もしくは2低く評価した事例がみられた。

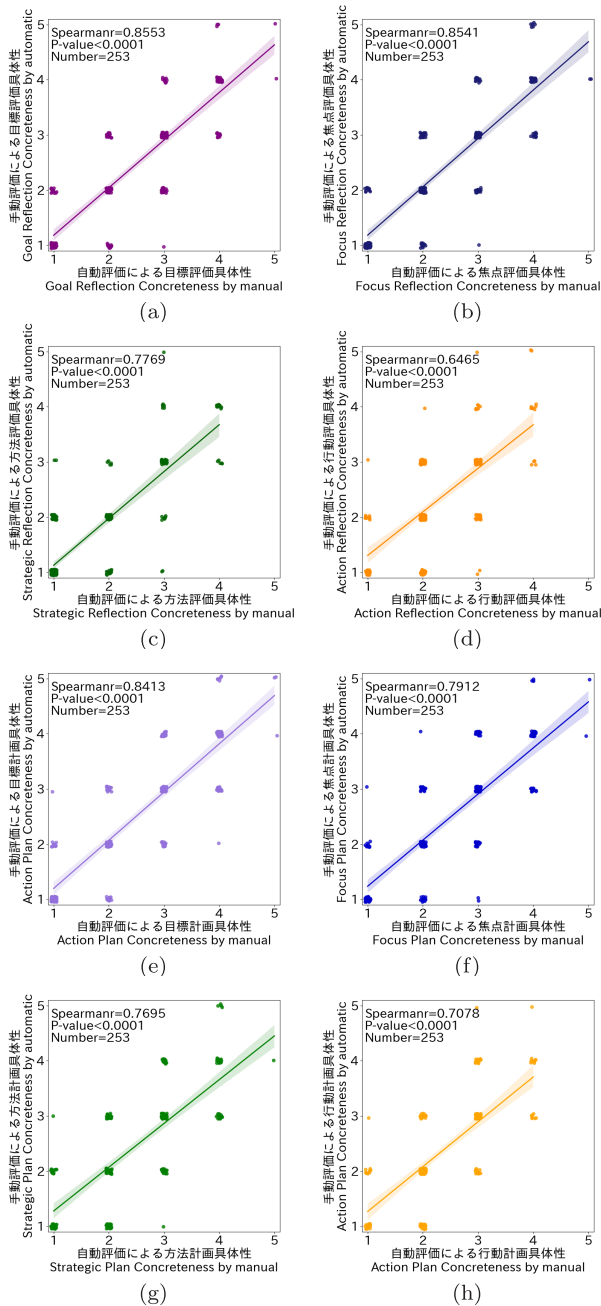


図8 項目ごとの具体性評価に関する手動評価と自動評価

Fig. 8 Item-specific concreteness evaluation by manual and automatic.

表6は、手動評価と自動評価の具体性評価が一致した記述事例の一部を示す。記述内容欄の括弧内に足場はずしLvを示した。Lv.3の事例は、自動評価モデルが、十分に記述内容があるため、自動評価による単語や文の重要な点と、それらの関係性の解釈がうまくできた可能性を示す。Lv.2では、加えてヒントによる語彙の類似性によってうまくできた可能性を示す。Lv.1の事例は、定形文による

表5 項目ごとの具体性評価誤差

Table 5 Item-specific concreteness evaluation error.

Evaluation items	-2	-1	0	1	2
目標評価具体性	0	41	168	43	1
焦点評価具体性	0	48	162	42	1
方略評価具体性	3	31	190	27	2
行動評価具体性	3	48	163	37	2
目標計画具体性	1	48	155	48	1
焦点計画具体性	2	49	149	51	2
方略計画具体性	1	53	136	62	1
行動計画具体性	2	46	161	44	0
Total	12	371	2294	354	10

表6 手動と自動の具体性評価が一致した記述内容

Table 6 Descriptions consistent with manual and automatic concreteness evaluation.

種類	記述内容（足場はずしレベル）	評価
1. 目標評価	前回の計画に対して順調に進めることができ、今回はディシジョンツリー上で文字設定とルール遷移設定を終わらせることができた。(Lv.3)	4
2. 焦点評価	本当にその選択肢はターゲットユーザーを想定して作成できているかを考えることが一番注目したことである (Lv.3)	4
3. 方略評価	条件について考えることでスムーズに学習用データの作成を進めることが出来た。(Lv.3)	4
4. 行動評価	計画したように行動できた。計画した行動で目標の課題を終えた。計画した行動は自分のスケジュールにあったと思う。(Lv.1)	3
5. 目標計画	まずは画像の表示設定を行っていきたいです。画像の表示設定が終われば次の課題、と段階を踏んでやっていきたいです。(Lv.3)	4
6. 焦点計画	言葉の表示設定をブラッシュアップしていくことに注目していきたい (Lv.2)	3
7. 方略計画	同じような素材でもそれぞれ違う角度から写真を比べた時に出てくる条件があるためそれをたくさん出し、データに反映していく (Lv.3)	4
8. 行動計画	いつどこでやるかのイメージができています。いつどこでやるかを調整しておく必要は今のところなさそう。今回の課題は負担がありそうに思う。(Lv.1)	2

類似性によってうまくできた可能性を示す。

表 7 は、手動評価と自動評価の具体性評価に誤差があった記述事例の一部を示す。これらの事例は、比較的文字数が少ない事例では、文や単語の重要な点、それらの関係性をうまく捉えられず誤差が生じた可能性を示す。誤字や文に違和感のあるものは、自動評価は教員が行うような文章の補完を行わなかった可能性を示す。授業内容との関係の認知期の違いなどの要因があった可能性を示す。事例ごとの分析は考察で示す。

表 8 の記述内容は、想定外の記述内容にもかかわらず、比較的うまく自動評価された事例を示す。課題内容と直接関係性が少ない記述が含まれるものの、学習者の工夫が含まれた記述内容である。手動評価とほぼ同等の自動評価がされた。

6.1.2 足場はずしレベルの自動設定

足場はずしレベルについて実験結果を示す。足場はずしレベルは、手動評価においても自動評価においても具体性評価値の合計を暫定閾値で区分することで自動設定され

表 7 手動と自動の具体性評価に誤差のあった記述内容

Table 7 Descriptions that had errors in the concreteness evaluation of manual and automatic.

種類	記述内容（足場はずしレベル）	手動	自動
1. 目標評価	公欠で休んでしまったので、取り返していきたい。(Lv.3)	1	3
2. 焦点評価	ディジョンツリー (Lv.3)	1	3
3. 方略評価	あまりうまくいかなかった。実行テストを行っても、言葉と出てくる文字の不一致がでてしまったため、何度もテストを行い修正を重ねるべきだと思う。(Lv.3)	5	3
4. 行動評価	テストでちゃんとできるようにする。(Lv.3)	1	3
5. 目標計画	Excelを終え、デッショントツリー作成に入る (Lv.3)	2	4
6. 焦点計画	著作権 (Lv.3)	3	1
7. 方略計画	以下に物を売るか (Lv.3)	1	3
8. 行動計画	毎日する事が大事である。学校、バイト終わりにする授業の空きコマなどを使いやってみる。(Lv.3)	5	3

表 8 想定外の記述内容への自動評価事例

Table 8 Automatic Evaluation of Unexpected Descriptions.

種類	記述内容	手動	自動
目標評価	3DCG でのロゴアニメーションを作って YouTube にアップロードしたものを設定した。	4	4
焦点評価	平面だったロゴを綺麗に立体化するにはどうするか。	4	3

た。図 9 からは、手動評価と自動評価の結果に強い正の相関+0.804 ($p<0.0001$, $n=253$) が読み取れる。図 9 の散布図は、縦軸が手動評価、横軸が自動評価による足場はずしレベル設定を示す。表 9 は、手動評価と自動評価によって自動選択された足場はずしレベルごとの件数を示した。足場はずしレベルごとの選択数にわずかな誤差が認められる。表 10 は、足場はずしレベルの誤差（自動評価の Lv.-手動評価の Lv.）の詳細を示す。自動評価が手動評価より低いレベルを設定した件数が 22 件あり、高いレベルを設定した件数が 23 件あった。

6.2 躓き度の自動評価

図 10 は、躓き度の評価について手動評価と自動評価の散布図を示す。縦軸が手動評価、横軸が自動評価、評価は躓き度が強いほど 5、弱いほど 1 を示す。手動評価と自動

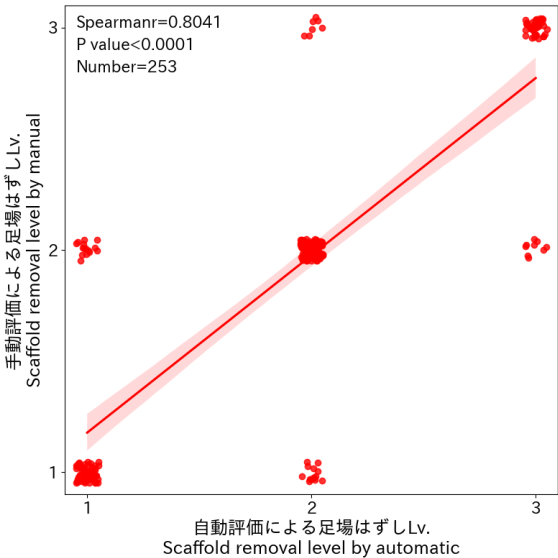


図 9 足場はずし Lv に関する手動評価と自動評価

Fig. 9 Manual and automatic evaluation of scaffold removal levels.

表 9 手動評価と自動評価に関する足場はずし Lv. 件数

Table 9 Number of manual and automatic evaluation scaffolds removed Lv.

Method of evaluation	Lv.1	Lv.2	Lv.3
手動 Manual	73	134	46
自動 Automatic	74	131	48

表 10 手動評価と自動評価に関する足場はずし Lv. 誤差

Table 10 Scaffolding Lv. error on manual and automatic evaluation.

Lv. by automatic - Lv. by manual	Count
-1	22
0	208
1	23

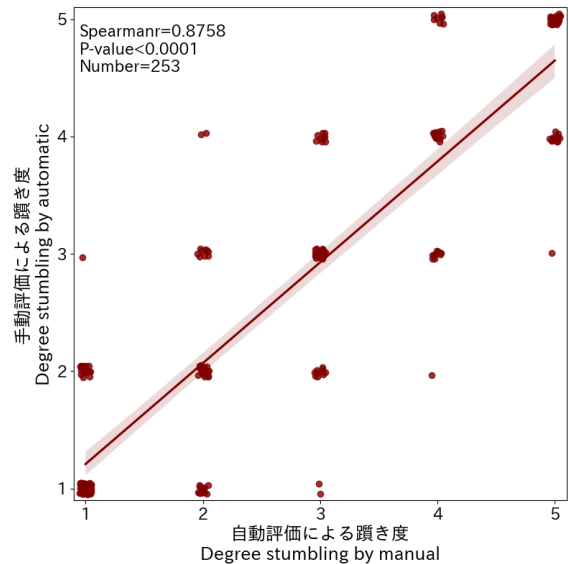


図 10 躓き度に関する手動評価と自動評価

Fig. 10 Manual and automatic evaluation of stumbling degrees.

評価の結果に強い正の相関+0.876 ($p<0.0001$, $n=253$) が示された。足場はずしレベルごとに分析すると、Lv.3 は正の相関+0.666 ($p<0.0001$, $n=46$)、Lv.2 は強い正の相関+0.8 ($p<0.0001$, $n=134$)、Lv.1 は正の相関+0.792 ($p<0.0001$, $n=73$) が示された。

表 11 は、躓き度評価値の誤差について詳細を示す。92 件の誤差±1 と 7 件の誤差±2 があった。手動評価の足場はずしレベル別で見ると Lv.2 (ある程度具体性がある記述) に誤差の件数が多い傾向が読み取れる。

表 12 は、手動と自動による躓き度評価が一致した記述例として、手動評価、自動評価とも躓き度 1 (躓きがほとんどみられない) の例を示す。この事例は、各項目について全体的に十分な文と単語が含まれていることから、自動評価は文や単語の重要な点、それらの関係性をうまく捉えることができていた可能性を示している。

表 13 は、手動と自動による躓き度評価に誤差があった記述例として、手動評価の躓き度 4 (だいぶ躓きがみられる)、自動評価の躓き度 2 (やや躓きがある) の誤差の例を示す。この事例は、各項目について文字数と文と単語の偏りがあることから、自動評価は文や単語の重要な点、それらの関係性をうまく捉えることができなかったために誤差が生じた可能性を示している。

6.3 足場はずしと記述内容

12 回目の記述内容から、足場はずし Lv. によって以下 3 つのテストデータの質の違いが読み取れる。10 回目の 128 件は Lv.3 のみであり、12 回目の 125 件は Lv.3 : 16 件、Lv.2 : 63 件、Lv.1 : 46 件が混在する。

(1) Lv.3 は、学習自身が吟味して工夫された記述されている。ただし、うまく記述できない学習者に短い文か

表 11 躓き度評価の誤差

Table 11 Difference in the degree of stumbling.

Difference	All Lv.	Lv.1	Lv.2	Lv.3
-2	3	0	3	0
-1	46	10	30	6
0	154	43	76	35
1	46	18	23	5
2	4	2	2	0

表 12 手動と自動による躓き度評価が一致した記述例

Table 12 Example of description with consistent manual and automatic stumbling degree evaluation.

記入項目	記述内容
目標評価	実際にディジションツリーが稼働するかや、条件の設定をした。
焦点評価	条件の仕組みを工夫することに 1 番注目した。
方略評価	かなりうまくできたので同じ方法、考え方で次回も乗り切りたい
行動評価	授業前やバイトの休憩時間
目標計画	今回の課題はかなり複雑な内容なので 3 つの段階ごとに分けて徐々に単純化していくやり方でいきたい。
焦点計画	今回の課題はかなり実践的な内容なので本番をイメージしながら調整していくと良いと思った。
方略計画	ただ取り組むのではなく、第三者目線なども考えて多面的に情報を汲み取る考え方で取り組みたい。
行動計画	授業の始まる前や休日、バイトの休憩時間

表 13 手動と自動による躓き度評価に誤差 (手動 4、自動 2) のあった記述例

Table 13 Examples of descriptions with errors in manual and automatic stumbling degree evaluations.

記入項目	記述内容
目標評価	課題用データをアップロードしようと思ったが、エクセルの行数が足りないことや、写真を変更しなければいけないことに気づいたので、あまり進むことはできなかった
焦点評価	どのように別のルートで画像を表示させるか
方略評価	あまりうまくいかなかった
行動評価	月・水・金の授業後
目標計画	まずは一週間の前半でデータのアップロードまで終了させ、後半で細かな調整を行う
焦点計画	どれだけ自分のイメージと近い演出にできるか
方略計画	授業資料と照らし合わせ、適切な方法で進める
行動計画	月・水・金の授業後

つ抽象的な記述がみられた。

- (2) Lv.2 は、提示されたヒントを参考に記述されている。
- (3) Lv.1 は、学習者が選ぶ選択肢に応じた定形文に沿って記述されている。

なお、Lv.1 の 46 件について、33 件に定形文に沿った記述がみられた。一方、残りの 13 件は定形文をあえて削除して書き直したとみられる記述があった。

また、うまく記述できない学習者については、短い文かつ抽象的な記述がみられた。具体的には「コツコツとやる」「ない」「思うままにやる」「空いてる時間」「特にない」などであった。これらは、Lv.3, Lv.2, および、Lv.1の定形文をあえて削除した記述にみられた。

以下は、12回目の記述内容の事例として計画の一部について、足場はずしLv.ごとの事例を示す。計画の事例を示した理由は、Lv.2において、ヒントの影響が分かりやすいためである。

Lv.3の事例は、自由記述欄に記入された事例を示した。

Lv.2の事例は、ヒントを活かして記述された事例を示した。提示されたヒントは「たとえば、まずは、デシジョンツリーの遷移を設定して、動作を確認してから、言葉の表示を設定していくとかかな。どうかな。自分なりのステップを作ってみよう。」であった。したがって、「ツリーを作成して」「言葉の位置を調整」の記述によって、ヒントを活かしていることが分かる事例である。

Lv.1は、定形文を編集して記述された事例を示した。「計画する目標地点」についての難易度、「自分の目標地点」について、自分が到達できると思う目標、「一番近い目標地点」について、穴埋め的に記述された事例である。

目標の計画

Lv.3 実行テストを繰り返し行い、改善点を見つける。デシジョンツリーの修正を行う。文字設定を行う。

Lv.2 ツリーを作成して、キャッチフレーズを紙に書き出してから、画像と言葉の位置を調整していく。

Lv.1 計画する目標地点は、困難。自分の目標地点は、初期設定までにある程度完成させる。一番近い目標地点は、動作確認までにある程度完成させる。

7. 考察

実験結果が示した自動評価と手動評価の比較において、以下3つの自動評価の結果が示された。

- (1) 具体性の自動評価 (目的 1). 順位相関係数+0.815
- (2) 躰き度の自動評価 (目的 1). 順位相関係数+0.876
- (3) 足場はずしレベルの自動設定 (目的 2). 順位相関係数+0.804

したがって、提案手法による自動評価用モデルは、本研究の目的1: 教員の手動評価を模倣して具体性と躰き度の自動評価ができていた可能性と、目的2: 自動評価で得た値を用いて足場はずしの自動設定ができていた可能性が示された。このことから、十分に実用に耐える性能と考えられる。以下に、実験結果について考察を加える。

7.1 自動評価の妥当性

まず、実験結果の全般において自動評価と手動評価に正の相関がみられた理由について考察を述べる。

自動評価は、教員をある程度模倣した評価ができていた

と考えられる。なぜなら、具体性の自動評価、躰き度の自動評価、足場はずしレベルの自動設定の3つとも、手動評価との順位相関係数に強い正の相関を示していたためである。また、表6は、付録A.1のループリックにある程度沿って教員が評価できていたと思われる事例であり、これらの評価が自動評価と手動評価で一致していたからである。ただし、教員の評価が付録A.1のループリックどおりにできていたかどうかの検証はできておらず、今後の課題としている。

(1) 標本が授業内容に依存していた。表6に示した1. 目標評価, 2. 焦点評価, 3. 方略評価, 5. 目標計画, 6. 焦点計画, 7. 方略計画の記述内容は、学習者の個人差よりも授業内容に依存した内容になりやすいと考えられる。そのため、交差検証用のトレーニングデータに他グループの授業内容に関する十分な標本が含まれたため、性能が高くなったと考えられる。

(2) 標本の範囲が定まっていた。交差検証用に用いたトレーニングデータは、自動評価の対象となる授業回10, 12回のみであった。そのため、標本の範囲が限定され評価値を付けやすくなり性能が高くなったと考えられる。なお、他の授業回やすべての授業回を用いた場合にも同様の結果が得られるのかについてはさらに検証を加える必要がある。また、10回, 12回目以外のトレーニングデータに含まれていない授業回について検証するなどトピック依存性の検証について今後の研究課題とする。

(3) 足場かけによる記述支援の影響。自動評価の精度を高めた要因として、12回目のテストデータに含まれた足場かけの影響が考えられる。なぜなら、12回目のテストデータは、足場かけとして記述支援を受けたデータを含むからである。12回目では、選択肢によって提示される定形文を用いた記述支援(足場はずしLv.1)が46件あった。そのうち33件に、ある程度定型文にあてはまるテキストデータがみられた。表6の「4. 行動評価」と「8. 行動計画」は、定型文によって、手動評価と自動評価が一致したとみられる事例である。

さらに、自由記述欄にヒントを提示した記述支援(足場はずしLv.2)も63件あった。与えられたヒントによって、記述内に使用される語彙の範囲がある程度限られた可能性が考えられる。表6の「6. 焦点計画」は、ヒントによって、手動評価と自動評価が一致したとみられる事例である。

一方、10回目のテストデータは、記述支援なしのデータであった。そのため、12回目のデータよりもうまく記述できない学習者のデータを多く含んでいた可能性が考えられる。10回目では、うまく記述できない学習者に短い文で抽象的な記述がみられた。したがって、足場かけの

あった12回目のテストデータは、足場かけがなかった10回目よりも自動評価の精度を高める要因が多く含まれていたと考えられる。なお、自動評価の性能評価について、記述時の足場は少しレベルごとの検証は行っていない。なぜなら、実際の運用では、足場は少しレベルが混在することを想定したからである。ただし、本研究によって当初の想定よりも、記述時の足場は少しレベルの違いによるデータへの質的影響があると考えられたため、足場は少しレベルごとの検証は今後の課題とする。

7.2 自動評価の誤り例とその分析

表13のように、自動評価と手動評価の躰き度に差異が生じる事例も確認された。そこで、誤差の事例について考察を加える。

7.2.1 具体性の自動評価における誤差の分析

具体性評価項目ごとの相関は、自動評価と手動評価は全体的に高い相関を示している。しかし、行動振り返り（行動評価）と行動計画の具体性評価は、他の具体性評価項目に比べて相関が低い傾向がみられた。その原因は以下の2つと考えられる。

- (1) 行動の個人差. 上述したように、目標、焦点、方略の記述内容は授業内容に依存する。しかし、行動は授業内容よりも学習者個別の都合に依存するため、より多くのトレーニングデータが必要と考えられる。トレーニングデータに十分な多様性が無く性能が低くなったと考えられる。
- (2) 高評価の標本の少なさ. 表3に示されたトレーニングデータに用いられた手動評価の件数を確認すると、行動について高い評価を付けている標本数が少ない。そのため、標本数が少なかった高い評価よりも、低い評価値を付けやすくなり、性能が低くなったと考えられる。

表7に示された誤差の理由について、ループリックに照らした考察を加える。以下の考察によって、授業内で用いられた言葉に対する認識や解釈の差、および、記述内容を補完した解釈から誤差が生じていた可能性が考えられる。

1. 目標評価 「公欠」「休んでしまった」「取り返したい」が学習に関係する内容かどうかの認識の違いによって誤差が生じたと考えられる。教員は、記述内容が学習内容に直接関係しないと判断したと考えられる。一方、自動評価は、何らかの振り返りで改善点が示されていると判断した可能性が考えられる。
2. 焦点評価 「ディジョンツリー」（デジョンツリーの間違い）について、自動評価は、学習に関する語彙であると誤認した可能性が考えられる。さらに、単語のみで評価3としたのは、トレーニングデータに単語のみで評価していた事例が含まれていた可能性が考えられる。一方、教員は記述に誤字があることに加え、単

語を示すのみでは振り返りとしては不足と判断したと考えられる。

3. 方略評価 記述の冒頭と、その後の記述の関係をどう認識して解釈したのかによって誤差が生じたと考えられる。自動評価は冒頭の適性を示した部分と、その後に続く具体的な部分との関係についてうまく評価できなかった可能性が考えられる。一方、教員は、文章の前後関係を補完して解釈したのち評価したと考えられる。
4. 行動評価 「テスト」が最終的な機会であることの認識の差によって誤差が生じたと考えられる。自動評価は（何度も実施されるような）「テスト」の機会に行うとした評価になったと考えられる。一方、教員は最終的な機会であることを認識しているため、行動計画の振り返りの記述として不十分であるとした可能性が考えられる。
5. 目標計画 「Excel」と「デジョンツリー作成」の関係の認知の違いで誤差が生じたと考えられる。教員は、「Excel」と「デジョンツリー作成」の2つが授業においては一体的な工程であることを知っている。一方、自動評価は、授業における2つの関係を認識していない。そのため、異なる2つの工程を段階的に進めることが記述されているとして評価した可能性が考えられる。
6. 焦点計画 「著作権」の扱いが異なることで誤差が生じたと考えられる。教員は、授業で扱った内容を含めて単語のみで具体的であるとした可能性が考えられる。一方、自動評価は、授業で扱った注意点であることを認識していない。そのため、用語のみでは、注意点について具体的ではないと判断した可能性が考えられる。
7. 方略計画 自動評価は、誤字を含め何らかの方法や考え方が記述されていると誤認した可能性が考えられる。一方、教員は、誤字を補完しつつも、課題を進めるための方法についての記述ではないと判断した評価と考えられる。
8. 行動計画 教員が文を補完して解釈し評価していたことで誤差が生じた可能性が考えられる。教員は「バイト終わりにする」と「授業の空きコマ」を分けて評価したと考えられるが、自動評価は、1つの具体的なタイミングがあるのみと評価した可能性が考えられる。

7.2.2 躰き度の自動評価における誤差の分析

自動評価された躰き度において、手動評価との誤差が比較的大きかった記述例について考察を加える。

表13の誤差のあった記述例では、学習者は計画に書いたことが実施できていないという経緯があった。しかし、振り返りでは、できているように記述していた。そのため、自動評価では躰き度2と評価された。一方、手動評価

した教員は、上記の経緯を把握していたため躰き度4とした可能性がある。つまり教員による手動評価では、学習者の経緯を加味して評価が行われたと考えられる。

このことから、教員の手動評価を模倣するためには、課題の制作状況や専用アプリの操作ログなどの経緯を含めたトレーニングデータが必要となる。このような学習者の状況や経緯を含めた評価手法は今後の研究課題とする。

表13について、ループリックに照らした考察を加える。以下の考察によって、自動評価は、教員とは異なる記述項目を重視して判断していた可能性が考えられる。つまり、躰き度の自動評価は、記述された項目のすべてが十分に具体的に記述されていない場合、少ない記述内容だけで、判断できるほど十分なトレーニングデータが無かった可能性が考えられる。

手動評価 振り返りでは、目標評価はある程度自己把握できているものの、焦点評価に結果との関連付けた振り返りがなく、方略評価、行動評価が具体的でない、計画では、目標計画は段階的に示されているものの、焦点計画、方略計画、行動計画が具体的ではないことから見通しに期待ができないと判断されたものと思われる。

自動評価 目標評価とある程度詳細な記述があり、あまりうまく進められていない記述があるものの、目標計画にある程度の過程が示されており、焦点計画、方略計画、行動計画にある程度の記述があることによって手動評価ほど躰いているとは判断されなかった可能性が考えられる。

7.3 想定外の記述内容について

一方、表8のように想定していない記述内容について、うまく評価されていた事例がみられた。事例は、授業課題のデジタルサイネージに、学習者自身の興味関心のある3DCGの映像を組み合わせた内容を記述していた。トレーニングデータに該当事例と似たような記述が含まれていないにもかかわらず、自動評価は手動評価とほぼ同様にうまく評価していた。

このように、授業内容に自身の興味関心を組み合わせる工夫は、自己調整学習において焦点の調整方略として歓迎すべき内容である。

また、この事例によってGPT-3を用いたことの妥当性が示されたと考える。なぜなら、「3DCGでのロゴアニメーションを作ってYouTubeにアップロードしたものを設定した。」は、「3DCG」「ロゴアニメーション」「YouTube」などの想定外の語彙を含む記述であっても、その文脈から複数の過程の記述であると判断された可能性が考えられるからである。

ただし、この事例から、記入項目や授業内容にまったく関係しない内容を具体的に記述した場合についても、具体

性を高く評価する可能性が考えられる。

なお、学習者個別の特性や興味関心に伴う焦点の調整方略と授業内容や記入項目との整合性の課題については今後の研究課題とする。

7.4 教員の過度な依存への対策

本研究の自動評価は、十分な性能が示された。しかし、本研究の自動評価に教員が過度に依存することは危険が伴う。なぜなら、上述の躰き度誤差分析から自動評価が学習者の躰きを見逃す事例が明らかになっている。そのため自動評価への過度な依存は、学習者の躰きを見逃す恐れがある。そこで、対策案としては、たとえば自動評価と教員による手動評価を授業回ごとに交互に運用する仕組みとすることによって、教員の負担を一定程度軽減しながら、自動評価への過度な依存を防ぐ対策が考えられる。

7.5 自動評価の誤差に対する学習者の許容範囲

具体性の自動評価の誤差は、足場はずしレベルの自動設定に影響して学習者が記述する内容に影響を与えると考えられるが、誤差による影響は学習者の許容範囲に収まると考えられる。なぜなら、以下2つの措置による。

- (1) 足場はずしLv.を3段階設けている。そのため、具体性評価8項目の合計値から設定される足場はずしLv.を高く見積もりされた場合に極端な足場はずしとなる事例を少なくできる。
- (2) 足場かけにおける記述支援でも記述内容を自由に編集できる機会を設けている。そのため、足場はずしLv.を低く見積もりされた場合に学習の工夫を記述する機会を得られる。

ただし、誤差が学習者に以下のような影響を与える可能性もある。

- (1) 足場はずしLv.が低く見積もられた場合、足場かけの記述支援が提示するヒントや定形文によってバイアスを受ける。そのため、学習者が主体的に個別の状況を吟味して工夫する機会を失う可能性が考えられる。
- (2) 足場はずしLv.が高く見積もられた場合、学習者は記述すべき内容が分からず、振り返りや計画をうまく記述できなくなる可能性が考えられる。

表11に示したような躰き度の誤差について、躰き度が低く見積もられた場合においては、学習者の許容範囲に収まらないと考えられる。なぜなら、教員などから学習支援を受ける機会を失う可能性が考えられる。そのため、今後の課題として、教員による確認を要とする仕組みの構築が必要となる。一方、躰き度が高く見積もられた場合については、学習者の許容範囲に収まると考えられる。なぜなら、不要な学習支援を受ける負担が若干あるものの、学習としては無駄にはならないものと考えられる。

8. まとめ

本研究が提案した2つの目的について、自動評価手法は、教員の手動評価をある程度模倣した自動評価ができ、十分に実用に耐える性能であったと考えられる。

- (1) GPT-3 を用いることで、具体性と躰き度について、教員の手動評価を模倣した自動評価ができる。
- (2) GPT-3 で自動評価した値を用いて足場はずしを自動的に設定できる。

ただし、本研究は実際の授業の一部から得られたトレーニングデータをそのまま使用した。そのために、標本数の少ない高評価値をうまく自動評価できない事例が生じていた。このことから、実際の授業データをトレーニングデータとして用いる場合、事前に評価値の分布を分析し、標本に偏りがある場合にはダミーの記述内容や評価を加えるなどの補完する措置が考えられる。

また、躰き度の自動評価では、教員の手動評価と異なり学習者の状況や経緯を考慮しなかった。そのために、学習者の躰き度を低く評価した事例が確認された。このことから、課題の制作状況や専用アプリの操作ログなどの経緯を含めたトレーニングデータを含めるなどの今後の研究課題が明らかになった。

さらに、足場かけによる記述支援が、自動評価の精度に影響を与えていたことが分かった。特に潜在的な支援の必要性を見つけたして学習者を支援するためには、精度よく自動評価できることが重要である。そのために、うまく記述できない学習者が学習状況を記述できる支援が重要である。そこで、今後の課題として、足場かけを改善する。具体的には学習状況の詳細な聞取りなどの自動対話の内容を改善する。

なお、自動評価への教員の過度な依存は、学習者の躰きを見逃す恐れがある。そのため、教員が自動評価に依存しない対策が必要である。自動評価は一定の性能を示したが、上述で示したようにうまく評価できない事例もある。

そもそも、各学生への個別最適化した学習支援を行うにはテキスト情報だけでは不十分であり、学習者の特性や興味関心などを考慮する必要がある。

GPT-3 を用いた本手法は、当然、それらの要素を考慮できていないので、個別最適化のためには人間の教員による評価と併用する必要がある。そこで、たとえば自動評価と教員の手動評価を交互に運用する仕組みにするなどの対策案が考えられる。また、自動評価への過度な依存を防止するためのインストラクションが不可欠となろう。

本研究が提案した自動評価手法は、実際の授業から蓄積したデータを活用することで評価を自動化する手法である。つまり、従来の学習支援において教員が行わなければならない記述内容の評価部分について、教員が行っていた評価を模倣して自動化する提案である。そのため、教

員が個別に学習支援に使用してきたアドバイスや足場かけ、足場はずしの手法と親和性が高いと考えられる。

なお、本研究から新たな研究課題として次のことが明らかになった。

- (1) 躰き度について学習者の状況や経緯を含めた評価手法。
- (2) 焦点の調整方略による想定外の記述と無関係の記述をうまく自動評価する方法。
- (3) 自動評価に対する教員の過度な依存への対策。
- (4) トレーニングデータに含む授業回、含まない授業回に対する自動評価の検証。
- (5) 振り返りと計画記入時の入力フォームの種類と自動評価の検証。

これらの研究課題についてさらなる検討と検証を行い、学習支援の自動化に向けた改善点を明らかにする。

謝辞 本研究は JSPS 科研費 (JP22K12325) および JST CREST (JPMJCR20D1) の助成を受けた。

参考文献

- [1] B. J. ジーマン, D. H. シャンク, 塚野州一: 自己調整学習の理論, 北大路書房 (2006).
- [2] 張 セイ, 森本康彦, 中村勝一, 宮寺庸造: 初歩の自己調整者の成長を促す自己調整モデルに基づいた支援システム, 電子情報通信学会技術研究報告, ET, 教育工学, Vol.111, No.473, pp.173-178 (2012).
- [3] Zimmerman, B. J.: Goal setting: A key proactive source of academic self-regulation, *Motivation and selfregulated learning*, Routledge, pp.267-295 (2012).
- [4] B. J. ジーマン, S. ボナー, R. コーバック: 学習スキルと自己効力感を高める 初版第3刷, 北大路書房 (2016).
- [5] 森本康彦: 次世代eポートフォリオシステムによる学習評価支援のための自己対話による学びの振り返り促進と成長の見える化モデル, 日本教育工学会第33回全国大会講演論文集, pp.157-158 (2017).
- [6] 合田美子, 山田政寛, 松田岳士, 加藤 浩, 齋藤 裕, 宮川裕之: 自己調整学習サイクルにおける計画とリフレクション: 授業外学習時間と英語力との関係から (教育実践研究論文<特集>1人1台端末時代の学習環境と学習支援), 日本教育工学会論文誌, Vol.38, No.3, pp.269-286 (オンライン), DOI: 10.15077/jjet.KJ00009649929 (2014).
- [7] Hirose, N., Shiramatsu, S. and Okuhara, S.: Development of chatbot to support student learning strategies in design education, *2021 IEEE International Conference on Agents (ICA)*, pp.1-6 (online), DOI: 10.1109/ICA54137.2021.00007 (2021).
- [8] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al.: Language models are few-shot learners, *Advances in neural information processing systems*, Vol.33, pp.1877-1901 (2020).
- [9] 網岡敬之, 森 裕生, 江木啓訓, 尾澤重知: 定量化した手書きワークシートを用いた学習評価の可能性の検討, 日本教育工学会論文誌, Vol.41, No.3, pp.245-253 (オンライン), DOI: 10.15077/jjet.41019 (2018).
- [10] 中山 実, 六浦光一, 山本洋雄: ノート記録の特徴情報とテスト得点との関係に関する一検討, 日本教育工学会

論文誌, Vol.39, No.Suppl, pp.53-56 (オンライン), DOI: 10.15077/jjet.S39038 (2016).

- [11] 丸山浩平, 森本康彦: カスタム機械学習モデルを用いた振り返り記述内容の分類とその活用方法の提案, JSiSE 研究会研究報告, Vol.33, No.5, pp.53-58 (2019).

付録

A.1 教員の手動評価におけるルーブリック

A.1.1 目標の振り返り具体性

- 評価 1: 目標が何だったかについて書かれていない
 評価 2: 遠い目標, あるいは, 抽象的な目標について書かれている
 評価 3: 具体的な目標について書かれている
 評価 4: 具体的な目標に対して, どのような過程を経たのかについて書かれている
 評価 5: 具体的な目標に対する過程でどこまで到達できたのか, できなかったのか, 対応などについて書かれている

A.1.2 焦点の振り返り具体性

- 評価 1: 注意点となったことについて書かれていない
 評価 2: 大まかな注意点について書かれている
 評価 3: 具体的な注意点について書かれている
 評価 4: 結果に寄与する注意点として書かれている
 評価 5: 注意点と成果や結果との関係などの説明が書かれている

A.1.3 方略の振り返り具体性

- 評価 1: 使用した方法や考え方について書かれていない
 評価 2: 使用した方法や考え方について抽象的に書かれている
 評価 3: 具体的に使用した方法や考え方について書かれている
 評価 4: 使用した方法や考え方の適性について書かれている
 評価 5: 使用した方法や考え方の適性と対策などの説明が書かれている

A.1.4 行動計画の振り返り具体性

- 評価 1: 取り組んだタイミングについて書かれていない
 評価 2: 取り組んだタイミングについて抽象的に書かれている
 評価 3: 取り組んだタイミングについて具体的に書かれている
 評価 4: 取り組んだタイミングの適性について書かれている
 評価 5: 取り組んだタイミングについて結果と対策などの説明が書かれている

A.1.5 目標の計画具体性

- 評価 1: 目標について書かれていない
 評価 2: 遠い目標, あるいは, 抽象的な目標について書かれている
 評価 3: 具体的な目標について書かれている
 評価 4: 目標に対して段階化した近い目標が書かれている
 評価 5: 段階化した目標について, 理由や達成見込みなどの説明について書かれている

A.1.6 焦点の計画具体性

- 評価 1: 注意点について書かれていない
 評価 2: 大まかな注意点について書かれている
 評価 3: 具体的な注意点について書かれている
 評価 4: 結果に寄与する注意点として書かれている
 評価 5: 注意点と見込まれる成果や結果との関係などの説明について書かれている

A.1.7 方略の計画具体性

- 評価 1: 使用する方法や考え方について書かれていない
 評価 2: 使用する方法や考え方についてか抽象的に書かれている
 評価 3: 具体的に使用する方法や考え方が書かれている
 評価 4: 使用する方法や考え方の適性が書かれている
 評価 5: 使用する方法や考え方の適性と, うまくいかないときの対策などの説明について書かれている

A.1.8 行動計画の計画具体性

- 評価 1: 取り組むタイミングについて書かれていない
 評価 2: 取り組むタイミングについて抽象的に書かれている
 評価 3: 取り組むタイミングについて具体的に書かれている
 評価 4: 取り組むタイミングの理由が書かれている
 評価 5: 取り組むタイミングが都合が悪いときの対策などの説明について書かれている

A.1.9 躓き度

- 躓き度 1: 目標, 焦点, 方略, 行動計画から状況把握ができており, 計画による見通しが十分期待できる.
 躓き度 2: 状況把握がほぼできており, 計画による見通しがある程度期待できる.
 躓き度 3: 状況把握ができていないものの, 計画による見通しはある程度できている.
 躓き度 4: 状況把握がある程度できているものの計画による見通しがあまり期待できない.
 躓き度 5: 振り返りで状況把握ができていない. 学習が進んでいない旨が書かれており, 計画による見通しに期待できていない.



廣瀬 伸行（正会員）

2005 中京大学大学院認知科学専攻修士課程修了，2007 愛知産業大学造形学部講師に従事，2020 から名古屋工業大学大学院工学研究科博士後期課程を履修中。



白松 俊（正会員）

2008 京都大学 情報学研究科 知能情報学専攻修了，2021 名古屋工業大学 大学院工学研究科情報工学専攻 教授。2018 から人工知能学会 市民共創知研究会 主査。シビックテックや市民協働の支援技術，市民参加型議論の支援技術の研究開発に従事。



奥原 俊（正会員）

2020 名古屋工業大学 大学院工学研究科 博士後期課程情報工学専攻修了，2021 京都大学 大学院情報学研究科 社会情報学専攻 特定助教，2022 三重大学 工学研究科 情報工学専攻・知能工学講座（人間情報学教育研究分野）講師。