英単語問題解答時の確信判定システムの精度検証

丸市 賢功 [1] 石丸 翔也 [2] Olivier Augereau[1] 黄瀬 浩一 [1]
Takanori Maruichi[1] Shoya Ishimaru[2] Olivier Augereau[1] Koichi Kise[1]

[1] 大阪府立大学大学院工学研究科
[2] ドイツ人工知能研究センター
[1] Graduate School of Engineering, Osaka Prefecture University
[2] German Research Center for Artificial Intelligence

Abstract: Feedback based on the combination of self-confidence and correctness of an answer can help learners identify misconceptions. In this study, we propose a self-confidence estimation method to aid the development of a system for vocabulary learning. The proposed method is based on the examination of eye movement and handwriting behaviors. To evaluate the performance of the proposed method, we examined the eye movement of 7 participants and the handwriting behavior of 8 participants while completing English spelling tests. The results revealed that our method was able to classify self-confidence into two classes with 87% accuracy in the typing dataset and 79% in the handwriting dataset.

1 はじめに

英単語の学習においては，定着度の確認を反復して行う必要がある．典型的な学習方法としては，英単語を暗記し，定着度を英単語学習ソフトなどのテストで確認し，その中から，間違えた英単語のみを復習対象として学習を行い，再度テストをして確認するというものがある．しかし，復習対象を明確にし，復習の優先順位を定めるには，解答した英単語に対する確信を考慮する必要がある．なぜなら，テストで正解した問題の中にも，確信が持てず偶然正解した英単語が含まれているからである．これらの英単語は次のテストでは間違える可能性が高く，復習対象とすべきである．また，テストで間違えた英単語については，正解する確信があった英単語も含まれている．これらの英単語は誤った暗記をしており重点的に復習する必要がある．

英単語に対する確信の有無を確認する方法には，確信の有無を逐次学習者が記録しておく方法があるが，面倒であり集中を阻害してしまう可能性がある．本研究では，近年の研究 [3, 4] から，学習者のテストをする際の振舞をもとに確信の有無を機械学習によって推定する．

本研究では，タイピングと筆記の 2 つの入力形式を扱う．タイピングについては，解答する際のタイピングの振舞を用いた手法が提案されている [5]．しかし，この手法は，タイピングへの慣れに大きく依存しておりタイピングに不慣れなユーザでは，確信がある場合と確信がない場合での振舞の差を識別することが難しい．そこで，全てのユーザにとってなじみのある筆記の振舞から確信を判定する手法を提案する．

また，提案手法を検証するために実験を行い，実験の精度，選択された特徴量についてタイピング手法との比較を行った．実験参加者はタイピング形式で 12 名，筆記形式で 11 名である．実験の結果，各手法の平均判定精度は，タイピング手法においては，ユーザ依存の推定では 87%，ユーザ非依存では 84%，筆記手法においては，ユーザ依存では 80%，ユーザ非依存では 75%であった．

なお，本研究は，大阪府立大学工学研究科倫理委員会の承認を得ている事を付記しておく．

2 関連研究

テスト時の振舞と問題に対する確信との関連は既に様々な研究で示されている．Wall らは，分数大小比較テスト解答時のユーザの確信と目の注視点のデータとの関連について調査している [6]．調査の結果，目の動きはテストの難易度によって大きく差がでることが示されたが確信判定には至っていない．Yamada らは，多肢選択式問題解答時の視点情報を用いて，解答している問題に対する確信度を判定する手法を提案した [7]．実験の結果，実験環境で 90.1%の識別率を達成しているが，検証は多肢選択式問題に限定されており，記述式問題における検証はなされていない．また，記述式問題については，浅井らが，タブレットで数学の問題に解答している際の筆圧や筆記速度，筆記間隔などの

情報を用いて，つまずいた箇所を検出する手法を提案している [1,9]．しかし，この研究も確信判定には至っていない．

Maruichi らは，スマートフォンでの英単語タイピング問題解答時のタイピング間隔を確信判定に用い，実験環境で 89.1%の識別率を達成している [5]．しかし，この研究はタイピングの振舞のみを対象としており，筆記の振舞を用いた確信判定の手法はこれまで提案されていない．

3 既存手法

本節では，タイピングまたは筆記の振舞を用いた確信判定の手法に関する詳細な説明を行う．提案手法はデータ取得，特徴量抽出，特徴量選択，識別境界の設定の 4 つのステップからなる．

3.1 データ取得

タイピングの手法では，キーを押す動作とキーを放す動作をキーイベントと定義する．キーロガーを用いてキーイベントの種類，タイムスタンプ，キーイベントが起こったキーの 3 種類の情報を取得する．

筆記の手法では，ペンがディスプレイに触れる動作とペンがディスプレイから離れる動作をペンイベントと定義する．ペンロガーを用いてペンイベントの種類，タイムスタンプ，ペンイベントが起こった点のディスプレイ上の x-y 座標，筆圧，消しゴムが使われたかどうかの 5 種類の情報を取得する．

3.2 特徴量抽出

取得したデータから，特徴量に変換する．本研究では，指またはペンがディスプレイに触れてから離れるまでの一連の動作を 1 ストロークと定義する．その定義をもとに解答時間，インターバル，スピード，消去比率の 4 種類の特徴量を抽出する．

(1) 解答時間： 解答時間とは，問題が提示されてからユーザが解答を終了するまでの時間間隔のことである．この値はストローク数が大きいほど，大きくなることが想定されるため，実際にはストローク数で割った値を特徴量として使用している．

(2) インターバル： 本研究では，ストロークとストロークの間にかかった時間をインターバルと定義する．すなわち，ユーザがディスプレイから指またはペンを放してから，次にディスプレイに触れるまでの時間間隔のことである．取得されたインターバルから問題ごとの平均，分散，最大値，最小値そして中央値を求めて用いる．なお，インターバルには解答開始前・終了後の時間間隔は含めないこととする．解答開始前の時間間隔とは，問題が提示されてからユーザが解答を開始するまでの時間間隔のことで，この値をインターバルに含めると，平均や最大値に影響を及ぼすからである．解答終了後の時間間隔とは，ユーザが最後のストロークを終えてから解答の決定ボタンを押すまでの時間間隔のことである．上記の特徴量に加え，解答開始前・終了後の時間間隔およびインターバルの総和を解答時間で割った比率を特徴量に含める．

表 1: 特徴量の一覧

No	特徴量
$f1$	解答時間 / ストローク数
$f2$-$f6$	インターバルの { 平均，分散，最大値，最小値，中央値 }
$f7$-$f8$	解答 { 開始前，終了後 } の時間間隔
$f9$-$f11$	{$f7$，$f8$，インターバルの合計 } / 解答時間
$f12$-$f16$	スピードの { 平均，分散，最大値，最小値，中央値 }
$f17$	消去回数 / ストローク数

(3) スピード： スピードについてはタイピングの手法と筆記の手法では若干定義が異なる．筆記の手法では，1 ストロークの長さを 1 ストロークにかかった時間で割った値なのに対し，タイピングの手法では，1 ストロークにかかった時間の逆数として定義している．これらの 1 ストロークのごとのスピードから問題ごとの平均，分散，最大値，最小値そして中央値を計算する．

(4) 消去比率： デリートキーまたは消しゴムを使った動作を消去と定義する．消去の回数は解答のの文字数に大きく依存するため，ストローク数での割った消去比率を使用することとする．

これらの特徴量は精度比較のため，タイピング・筆記双方において共通して取得できる情報を使用している．筆記でのみ取得できる筆圧は使用していない．したがって，本手法は圧力センサを搭載していない廉価版のタッチディスプレイやタッチペンにも適用することが可能である．

3.3 特徴量選択

計算された特徴量について，有効な特徴量を選択する．本手法では Forward stepwise selection を用いる．その上で，判定精度が最も高い特徴量の組合せを選択する．

3.4 識別境界の設定

選択された特徴量をもとに，SVM を用いて識別境界の設定を行い，確信あり（ラベルは 1 とする）・確信なし（ラベルは 0 とする）の 2 クラスに分類する．

4　実験

提案手法の有効性を示すための実験を行った．本章では，実験参加者や実験手順，精度の評価方法について簡単な説明を行う．

4.1　実験設定

4.1.1　タイピング

タイピング形式については 12 名（男性 9 名，女性 3 名）の日本の大学に通う日本人大学生に，図 1 (a) のような形式の問題を解くよう依頼した．問題にはリクルート社の英単語アプリ，スタディサプリ英単語の問題を使用した．

実験手順について述べる．まず，10 問の英単語問題を解答してもらい，10 問の解答終了後に，各々の解答について確信度を (1) 確信あり (2) 確信なしの 2 通りで報告するよう指示した．解答が分からない場合には問題をスキップすることも可能であり，その問題については推定には用いないこととした．この手順を合計 12 回繰り返し，解答する際のデータを記録した．

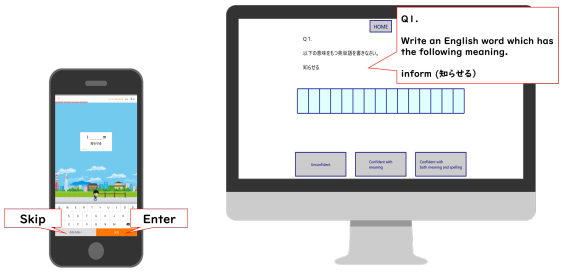
4.1.2　筆記

筆記形式については 11 名（男性 9 名，女性 2 名）のドイツの大学に通う日本人または中国人大学生に，図 1 (b) のような形式の問題を解くよう依頼した．問題には TOEIC テストでよく問われる英単語を出題した．

実験手順について述べる．まず，10 問の英単語問題を解答してもらい，1 問の解答終了ごとに確信度を (1) 意味・スペルともに確信あり (2) 意味にのみ確信あり (3) 意味・スペル共に確信なしの 3 通りで報告するよう指示した．しかし，(2) に対応するデータが少なすぎたため，推定時には (2) と (3) を統合し，(1) を確信あり，(2) または (3) を確信なしとした 2 値分類を行うこととした．解答は小文字ブロック体で，インターバルがとりやすいように 1 マスに 1 文字を入れるような形で解答するよう指示した．解答が分からない場合には問題をスキップすることも可能であり，その問題については推定には用いないこととした．この手順を合計 30 回繰り返し，解答する際のデータを記録した．

4.2　評価方法

精度の評価方法については，ユーザ依存，ユーザ非依存の 2 つの場合について検証を行う．ユーザ依存とは，識別器の学習に際して，特定のユーザのデータのみを用い，同一ユーザに対して確信判定を行う検証法をいう．ユーザ依存で学習された識別器を他のユーザに用



(a) タイピング　　(b) 筆記

図 1: 実験フォーマット

いることはできない．ユーザ依存での学習を行うことで，そのユーザが解答する際に確信がある場合あるいは確信がない場合にどのような振舞いを行うのかを学習することができる．その振舞いの差が識別できれば，確信判定がより容易になる．評価方法については，各実験参加者について，データを 10 のセクションに分割し，1 セクション分をテストデータ，残りのセクションを学習データとして分割し，全てのセクションについて評価を行う Leave-one-section-out Cross-validation を用いた．

ユーザ非依存とは，学習データを取得したユーザとは異なるユーザに対して識別器を適用する検証法である．ユーザ非依存で確信の判定が可能であれば，まだ一度も問題を解いたことのない新しいユーザの確信を判定する場合でも，既にあるデータセットを用いて判定を行うことができる．したがって，実用上はユーザ非依存での判定精度も高ければよりよい手法であるといえる．したがって，ユーザ非依存の場合についても検証を行っておく必要がある．評価方法については，ユーザ 1 人分データをテストデータ，残りのユーザのデータを学習データとして，全てのユーザについて評価を行う Leave-one-user-out Cross-validation を用いた．

5　結果と考察

5.1　識別率

実験から得られた識別率を表 3 に示す．ここでの Baseline とは，ユーザアンケートの際の確信ありとラベル付けされた問題の割合である．各実験参加者ごとの識別率は図 2 に示す通りである．

提案手法の識別率が Baseline と比較して統計的に有意であるかを符号検定 [8] を用いて確認した．その結果を図 2 内の*印で示す．タイピング手法では，ユーザ依存，ユーザ非依存ともに Baseline との有意な差が確認できた．一方，筆記手法では，ユーザ依存のみ Baseline

表 2: データセットの詳細

(a) タイピング

実験参加者	問題数	確信ありの割合
T1	93	0.78
T2	93	0.76
T3	103	0.49
T4	71	0.48
T5	96	0.50
T6	87	0.74
T7	96	0.67
T8	90	0.62
T9	94	0.51
T10	107	0.68
T11	75	0.61
T12	64	0.66
Ave.	88.7	0.61
Wt. Ave.		0.61

(b) 筆記

実験参加者	問題数	確信ありの割合
H1	209	0.50
H2	134	0.54
H3	134	0.37
H4	135	0.77
H5	51	0.67
H6	153	0.59
H7	141	0.65
H8	208	0.79
H9	225	0.85
H10	178	0.62
H11	263	0.75
Ave.	166.5	0.64
Wt. Ave.		0.66

表 3: 実験結果

(a) タイピング

評価方法	識別率	AUC	
		確信あり	確信なし
baseline	0.63 ± 0.11	0.63	0.37
ユーザ依存	0.87 ± 0.04	0.91	0.87
ユーザ非依存	0.84 ± 0.05	0.87	0.79

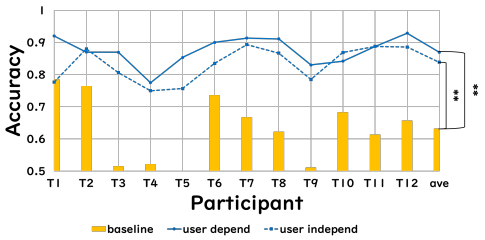
(b) 筆記

評価方法	識別率	AUC	
		確信あり	確信なし
baseline	0.67 ± 0.10	0.67	0.33
ユーザ依存	0.80 ± 0.05	0.87	0.73
ユーザ非依存	0.75 ± 0.07	0.87	0.64

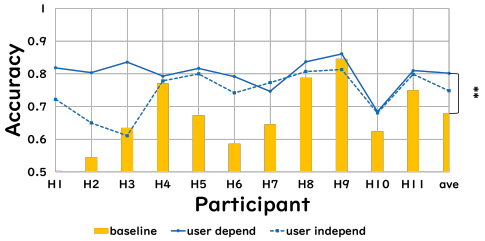
との有意な差が確認できた．この結果から，筆記の振舞はタイピングの振舞に比べて確信の識別が難しいことが分かった．

ユーザ非依存の識別率がユーザ依存の場合によりも低くなる理由としては，2 つの要因が考えられる．1 つには，一部のユーザには他のユーザには見られない固有の振舞をすることがあり，その振舞はユーザ非依存では学習できないということと，もう一つには，それぞれのユーザについて，確信の有無の判断基準にはばらつきがあり，ユーザ非依存での判定にはその判断基準が反映されないということである．

一方，一部のユーザでは，ユーザ依存の識別率がユーザ非依存の場合によりも低くなるという現象もみられた．この理由としては，ユーザ非依存の場合には，学習サンプル数はユーザ依存の場合に比べて多くなるので，より多くの振舞を学習ことができ識別率の向上が見られたと考える．



(a) タイピング



(b) 筆記

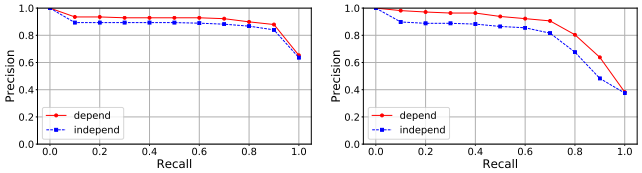
図 2: 各実験参加者ごとの識別率：*印は提案手法での識別率が Baseline と比較して統計的に優位であることを示している ($p < 0.01$)．

5.2 Precision-Recall 曲線

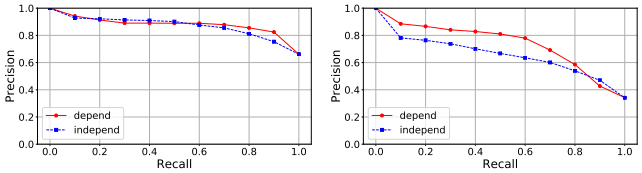
識別結果から得られた 11 点 Precision-Recall 曲線 [2] を確信あり，確信なし双方の場合について描画した（図 3 参照）．AUC (Area Under the Curve) については，比較のために，表 3 に示している．タイピング・筆記双方について，確信なしの場合の AUC が確信ありの場合よりも低いことが確認された．この理由としては，どちらのデータセットにおいても確信なしち回答された問題の割合が低かったからであると考える．これにより，確信なしの場合の振舞が十分に学習されず，識別器は確信ありと推定しやすくなっていると考えられる．

5.3 選択された特徴量

各特徴量と確信の有無の Pearson 相関係数を図 4 に示す．赤くプロットした点が選択された特徴量である．図 4 から，インターバルに関する特徴量は確信の有無と負の相関があることが分かる．すなわち，確信がない場合は，インターバルが長くなるということである．消去比率についても同様に，多ければ多いほど確信がないという傾向がみられた．したがって，これらの特徴量は確信判定に有効であるといえる．また，全特徴量について確信の有無と強い相関を持った特徴量は確認できなかった．そのため，選択された特徴量一つ一つの組み合わせが確信判定に寄与しているのではないかと考える．

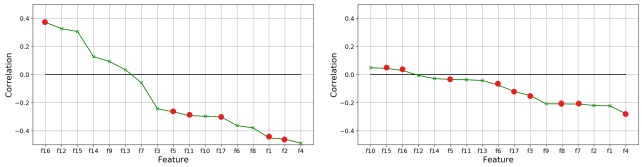


(a) タイピング，確信ありの場合　(b) タイピング，確信なしの場合



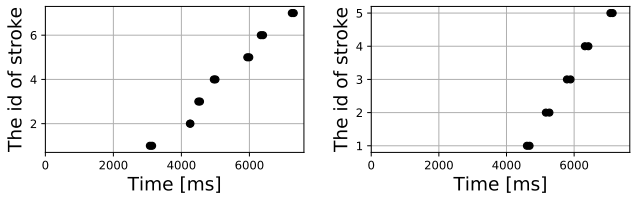
(c) 筆記，確信ありの場合　(d) 筆記，確信なしの場合

図 3: 11 点 Precision-Recall 曲線



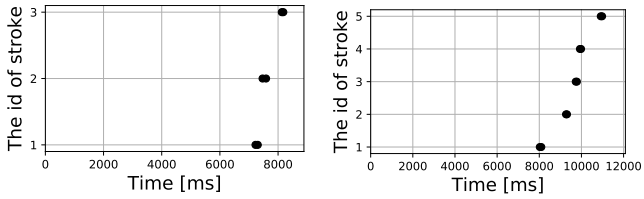
(a) タイピング　(b) 筆記

図 4: ユーザ非依存の推定における各特徴量と確信の Pearson 相関係数：赤くプロットした点は選択された特徴量



(a) True positive

(b) False positive



(c) False negative

(d) True negative

図 5: 各識別パターンにおけるタイピングの振舞の例：赤く示した文字はユーザが実際にタイプしたもの

5.4　誤識別の例

図 5，6 は各識別パターンにおけるタイピング及び筆記の振舞の例である．横軸は問題が提示されてからユーザが解答を終了するまでの時間を示しており，縦軸はストロークが行われた回数を示している．図中に示されている不連続な直線は 1 ストロークを表す．ここで，“Positive” とは確信あり，“Negative” とは確信なしのことである．例として，“False positive” とはユーザアンケートにおいて確信なしと回答されたにもかかわらず，識別器によって確信ありと判定された識別パターンのことを指す．各図の下部に示されている文字は実験参加者の実際の解答である．

5.4.1　タイピング

図 5 (b) は False positive の識別パターンの一例である．このサンプルでは，比較的短いインターバルがコンスタントに確認できる．理由としては，この文字列がキー同士の距離が短く短時間でタイプできるものだったことが考えられる（“o”, “p”, “p” など）．このような場合は，提案手法は確信があると誤識別しやすいことが分かる．一方，図 5 (c) のような False negative のパターンについては，ユーザが解答を始めるまでに時間が長くかかりすぎたため，解答時間（$f1$）の値が大きくなり，確信がないと識別されたことが推測される．

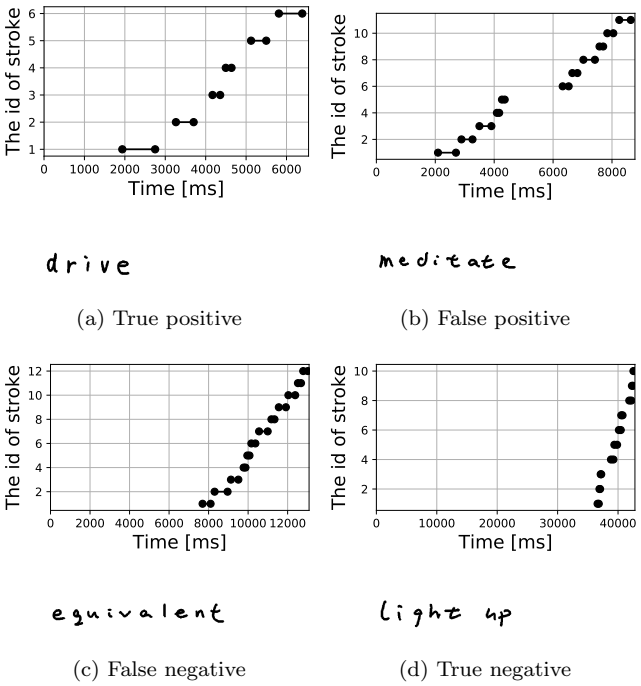
5.4.2　筆記

筆記の振舞についても，タイピングと類似の傾向が見られた（図 6 (b) 参照）．筆記の振舞が図 6 (a) のパターンに似通っているため，判定に失敗したことが考えられる．図 6 (c) のパターンでも，ユーザが解答を始めるまでに時間が長くかかりすぎたため，解答開始前の時間間隔（$f7$）に影響が出て誤識別を引き起こしたと考察する．

6　まとめ・今後の課題

本稿では，より効率的な英単語の復習を実現するために確信をタイピングまたは筆記の振舞を用いて判定する手法を提案した．実験の結果，各手法の平均判定精度は，タイピング手法においては，ユーザ依存の推定では 87%，ユーザ非依存では 84%，筆記手法においては，ユーザ依存では 80%，ユーザ非依存では 75%であった．

今後の課題としては以下の 3 点があげられる：



(a) True positive (b) False positive

(c) False negative (d) True negative

図 6: 各識別パターンにおける筆記の振舞の例 : 文字はユーザが実際に書いたもの

1. 確信判定がしやすい人としにくい人を比較し，その振舞について分析すること.
2. 学習効果検証のための実験を実施し確信フィードバックが本当に復習に有効であるかを確認する.
3. 確信のフィードバックによって学習効果が現れる人と現れない人を比較し，その振舞について分析すること.

謝辞

本研究の一部は、JST CREST (Grant No. JPMJCR16E1) の補助による.

参考文献

[1] Hiroki Asai and Hayato Yamana. Detecting student frustration based on handwriting behavior. In *Proceedings of the Adjunct Publication of the 26th Annual ACM Symposium on User Interface Software and Technology*, pages 77–78. ACM, 2013.

[2] Ricardo Baeza-Yates, Berthier de Araújo Neto Ribeiro, et al. *Modern information retrieval*. New York: ACM Press; Harlow, England: Addison-Wesley,, 2011.

[3] Andreas Dengel. Digital co-creation and augmented learning. In *Proceedings of the The 11th International Knowledge Management in Organizations Conference on The changing face of Knowledge Management Impacting Society*, page 3. ACM, 2016.

[4] Shoya Ishimaru, Soumy Jacob, Apurba Roy, Syed Saqib Bukhari, Carina Heisel, Nicolas Großmann, Michael Thees, Jochen Kuhn, and Andreas Dengel. Cognitive state measurement on learning materials by utilizing eye tracker and thermal camera. In *2017 14th IAPR International Conference on Document Analysis and Recognition*, volume 8, pages 32–36. IEEE, 2017.

[5] Takanori Maruichi, Koichi Kise, Olivier Augereau, and Motoi Iwata. Keystrokes tell you how confident you are: An application to vocabulary acquisition. In *Proceedings of the 2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2018 ACM International Symposium on Wearable Computers*, pages 154–157. ACM, 2018.

[6] Jenna Wall, Clarissa Thompson, and Bradley J Morris. Confidence judgments and eye fixations reveal adults' fractions knowledge. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, pages 2571–2576, 2015.

[7] Kento Yamada, Koichi Kise, and Olivier Augereau. Estimation of confidence based on eye gaze: an application to multiple-choice questions. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pages 217–220. ACM, 2017.

[8] 高木 英行. 使える!統計検定・機械学習-i : 2 群間の有意差検定. システム／制御／情報, 58(8):345–351, 2014.

[9] 苑田翔吾 山名早人 浅井洋樹, 野澤明里. オンライン手書きデータを用いた学習者のつまずき検出. In *DEIM Forum*, volume 2012, 2012.