# クラウドワーカの品質改善における参考回答提示の短期的・長期的効果

筑波大学
図書館情報メディア研究科
2019年03月
小林　正樹

# 目次

# 図目次

# 第 1 章

# 序論

クラウドソーシングは，人間の作業と計算機ネットワークによる情報処理を組み合わせることで様々な問題に取り組む手法である．作業の依頼者であるリクエスタが，不特定多数の作業者であるワーカに対して作業であるタスクを依頼するのが基本的な枠組みである．本稿では，画像や映像，音声へのタグ付けや分類，文章校正などの作業を扱うマイクロタスク型クラウドソーシングに注目する．

クラウドソーシングにおいて，成果物の品質を保証することが重要な研究課題の 1 つである．成果物の品質が低くなる要因としては，人間による作業が伴うことから成果物の一部に誤答が含まれる可能性や，ランダムな回答により単に報酬を受け取ることを目的とするスパムワーカの存在が挙げられる．これまでに多くの研究がこの問題に取り組んでおり，マイクロタスク型クラウドソーシングの大部分を占めると考えられる分類タスクやタグ付けタスクでは，主に次の 3 つのアプローチが用いられる．

1 つ目は優れたワーカを発見し，彼らに対してタスクを割り当てる方法である．例えば，Amazon Mechanical Turk ではリクエスタからの評価が高いワーカに対して作業を割り当てる MTurk Master Worker と呼ばれる仕組み[*1]を利用することが出来る．2 つ目は，同じタスクを複数のワーカに割り当て，複数のタスク結果を集約することである．最も単純な方法としては多数決が挙げられるが，ワーカやタスクの性質を考慮した様々な手法が提案されている．3 つ目は，個々のワーカからより良い結果を引き出す方法である．Shah らは，ワーカがタスクに回答したした後に，同様のタスクに回答した別のワーカの回答を提示し，回答を訂正する機会を与える自己補正と呼ばれる手法を提案した．自己補正はリクエスタがタスク画面を編集できる機能を持つ一般的なクラウドソーシングプラットフォームにおいて，タスクに適用が可能である．

Shah らは，ワーカのステージ 1 での成績が低い場合に，特に自己補正が有効であると主張した．ここで重要なのは，ワーカがステージ 1 での誤りに気づくことが出来た場合に，ステージ 2 においてステージ 1 の誤答を訂正出来ることである．つまり，タスクに自己補正を導入することで，ワーカに彼ら自身の誤りに気づかせる機会を与えることが出来るのである．

しかし，自己補正が提案された論文では，自己補正の有効性についてシミュレーションによる評価のみが行われており，現実のクラウドソーシング環境においても同様の効果が得られるかは不明である．実際のクラウドソーシング環境においても自己補正が有効であるかは興味深い課題である．

さらに，自己補正がワーカに対して無自覚な学習をもたらすことが出来るかについても注目すべき点である．ワーカが自己補正タスクに取り組むことにより，その後の作業の品質を改善

[*1] https://www.mturk.com/worker/help#what_is_master_worker

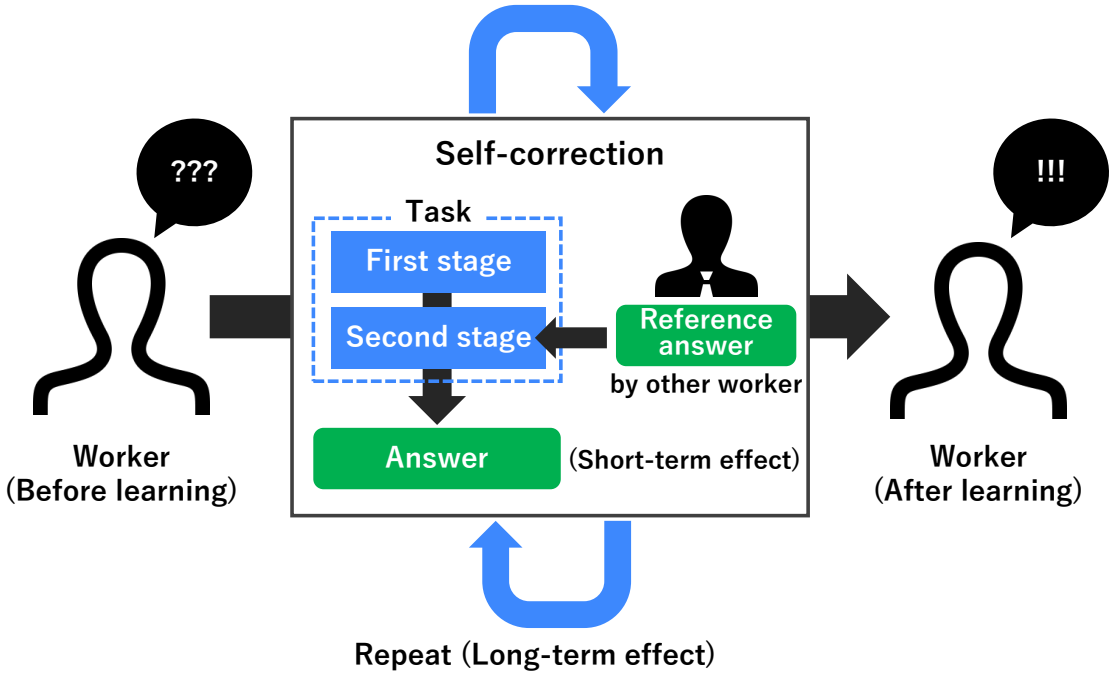


図 1.1　本研究の概要

できるならば，正答が既知の課題を利用した訓練フェーズを導入することなく，ワーカの品質を改善できることを意味する．視覚的な作業を行う能力は，繰り返すことにより，すなわち知覚的な学習によりその速度や精度が改善されることが知られている [1]．知覚学習は意図せず無自覚に生じるものであり [2]，複数の研究が，視覚的な分類課題における知覚学習を報告している [3]．したがって，自己補正を繰り返すことにより，視覚的な分類タスクを行う作業者の分類精度が向上すると考えられる．

本研究では，現実のクラウドソーシング環境における自己補正の短期的および長期的効果を検証するための実験結果について報告する（図 1.1）．

本研究における主な実験結果は次のとおりである．

1. 自己補正によりタスク結果の品質改善（以降では自己補正の短期的効果と呼ぶ）は，現実のクラウドソーシング環境においても見られた．
2. 自己補正の第 2 段階で提示する参考回答は重要な要因であることが示された．さらに，より信頼性の高い参考回答を提示することで自己補正がもたらす効果を増大させることが示唆された．
3. ワーカが自己補正を繰り返すことにより，ワーカ自身の回答品質が改善された（以降では自己補正の長期的効果と呼ぶ）．この長期的な効果は，ワーカが同様のタスクに繰り返し取り組むことによって生じた知覚学習であると説明できる．この結果から，自己補正がもたらす効果が，以降の同様のタスクについても良い影響を与えることが示唆された．
4. 自己補正の長期的効果が，学習に用いた課題とは異なる課題においても正答率の改善をもたらすかを明らかにするために，自己補正を適用した学習タスクと評価タスクで異なる課題を用いる実験を検討した．その結果，学習の転移は見られなかった．

本研究の各実験では，成果物の品質にかかわらずワーカに対して定額の報酬を支払った．一

方で，Shah[4] らは，自己補正に適用可能な報酬設定の手法を提案している．にもかかわらず，本研究で示す結果は Shah[4] の主張を支持するものであることから，自己補正は任意の報酬設定の手法が適用出来ないような環境においても有効であることが示唆された．Shah[4] が提案した手法を始めとする，様々な報酬設定の手法を現実にクラウドソーシングにて検証することは注目すべき点の 1 つである．

実験 1A では自己補正の短期的な効果と長期的な効果を鳥の画像分類課題で検討した．実験 1B では，実験 1A で見られた効果が，別の難易度の高いタスクにおいても同様の傾向であるかを，絵画の分類課題を用いて検討した．実験 2 では，実験 1A および実験 1B で見られた自己補正の長期的効果が，類似した別の課題においても品質の改善をもたらすかを，類似した複数の画像分類課題を組み合わせて検討した．

# 第 2 章

# 関連研究

クラウドソーシングにおいて，成果物の品質を保証することは重要な研究課題の 1 つであり，これまでに様々な研究がこの問題に取り組んできた [5].

主要な戦略の 1 つは，同じタスクを複数のワーカに割り当て，得られた複数の結果を集約して最終的な結果とする方法である．多数決は最も単純であり広く用いられている手法である．クラウドソーシングの文脈においては，ワーカの質は，得られた回答の傾向，回答のクラスタリング結果の利用など，より信頼性の高い集約結果を得るための手法が提案されている [6] [7] [8]．別の戦略としてはタスク設計の改善や [9]，タスク結果に基づく報酬決定手法などが挙げられる [10] [11]．Shah らの自己補正はこれらの手法と組み合わせて利用することも可能である．

ワーカへのフィードバックに着目したさまざまな研究があり，フィードバックによってワーカから得られるタスク結果の品質が向上することが知られている．Revolt [12] や Microtalk [13] ではあるワーカの回答を別のワーカが評価し，その評価を確認した上で回答を変更する機会を与える仕組みが用いられている．Shepherd はワーカの自己評価と様々な形態の外部評価を組み合わせるクラウドソーシングのためのフィードバックシステムである [14]．Shah らの自己補正では同じタスクに回答した他者の回答を提示するという単純なフィードバックを用いる．しかし，このフィードバックがどのように機能するかは明らかでない．

ワーカによるワーカ自らの評価には偏りがあると知られている．Gadiraju らはクラウドワーカが彼らの実際の能力についての認識に欠けていることが多いことを示した [15]．このようなバイアスを自己補正の枠組みに取り入れることは，自己補正における興味深い課題の 1 つである．

ワーカから得られるタスク結果を改善するために，ワーカの回答精度の改善に注目する場合，ワーカに対して本番のタスクを割り当てる前に訓練タスクを割り当てる手法が広く用いられている．ワーカに訓練タスクを割り当てた後に本番タスクを割り当てることで，本番タスクの品質が改善されることが知られている [16]．このような手法では，リクエスタやクラウドソーシングプラットフォーム運営者が訓練のためのタスクを用意する必要がある．また，ワーカに対して正誤のフィードバックを与える場合にはタスクと対応する正解を事前に得ておく必要がある．鈴木らは，ワーカが作業に必要なスキルを獲得することを支援するために，ワーカに対してインターンとメンターという関係を設けるマイクロインターンシップの仕組みを提案した [17].

マイクロタスクにおける知覚学習には，ワーカへのフィードバックが重要であると考えられる．Abad らは，誤った回答をしたワーカに対してルールに基づいてフィードバックを与えることが，ワーカの訓練に効果的であることを示した [18]．本研究における関心は，この様なワーカの知覚学習が，自己補正で提示するような単純なフィードバックでも生じるかである．

本研究では，ワーカが自己補正を適用したタスクをこなす過程で，ワーカの知覚学習が観察され，ワーカから得られる回答の品質が改善されるかどうかを検証する．知覚学習が生じるための重要な要素として，ワーカが同じ作業を繰り返して行うことが挙げられる．Law らは，ワーカが同じタスクを長時間こなすことを促すためのインセンティブ設計について議論した [19]．自己補正の枠組みにこのような仕組みを導入することは興味深い課題の 1 つである．

本研究の実験では，自己補正にて信頼性の高いワーカから得られた参考回答を提示する場合に，作業に取り組むワーカから得られたタスク結果の品質が改善されるかを明らかにする．ただし，自己補正の第 2 段階において信頼性の高い回答を提示するための方法は，自明でない．信頼性の高いワーカの回答を提示するために重要となるのが，ワーカの品質を評価する仕組みである．ワーカの品質を評価する仕組みやアルゴリズムについては様々な研究がなされている [20] [21] [22]．最も単純な方法は，ワーカに割り当てるタスクの中に，ワーカの能力を測定するための特別なタスクを追加することである．クラウドソーシングの文脈ではゴールドスタンダードクエスチョンと呼ばれている．より正確にワーカの品質を推定するために，ワーカが作業を介した直後の数タスクによりワーカの品質推定を行うのではなく，作業の中盤や後半においても継続的にゴールドスタンダードクエスチョンを割り当てることが効果的であると示されている [23]．さらに，ワーカの評価のためにゴールドスタンダードクエスチョンを使用せず，複数のワーカの回答の照合結果からワーカの品質を推定する手法も提案されている [24] [25]．クラウドソーシングでは正解が未知の課題を扱うことが多いことから，これらは有効な手段であると考えられる．

# 第 3 章

# 本研究のアプローチ

## 3.1　自己補正

この節では，Shah[4] らが提案した自己補正について，彼らの論文の貢献を説明する．

### 3.1.1　タスクの構成

一般的なクラウドソーシングサービスでは，ワーカは自身の誤りを発見して訂正する機会がない．しかし，多くのワーカ（スパムワーカなどを含まない）においては，誤りに気づく機会を提供することによって，ワーカが自らの回答を訂正することが出来ると考えられる．自己補正は，クラウドワーカらの成果物の品質を高めるためにのタスク設計である．自己補正では，ワーカは同じ質問に対して 2 回回答する機会が与えられる．1 回目は，通常のクラウドソーシングタスクと同様に回答し，2 回目では他者の回答を照らし合わせて回答を変更することが出来る．

### 3.1.2　報酬アルゴリズム

自己補正を適用したタスクでは，第 2 段階で他者の回答を考慮するのではなく，単に自身の回答を他者の回答で置き換えてしまうようなワーカの存在が想定される．そこで，Shah らは自己補正のための報酬アルゴリズムを提案した．彼らのアルゴリズムは，第 1 タスクで正答すると最も価値が高く，第 2 段階で他者の回答を支持すると低くなるような設定となっている．

### 3.1.3　シミュレーション

Shah らは，自己補正の有効性を明らかにするための，シミュレーションによる実験を行った．シミュレーションでは，自己補正を適用したタスクと通常のタスクを比較した．シミュレーションの結果は，自己補正を適用したタスクのほうが，最終的に得られる成果物の品質が高くなるというものである．彼らによれば，自己補正を適用することにより，成果物を用いるアプリケーション（例えば機械学習など）の品質が改善されるという．

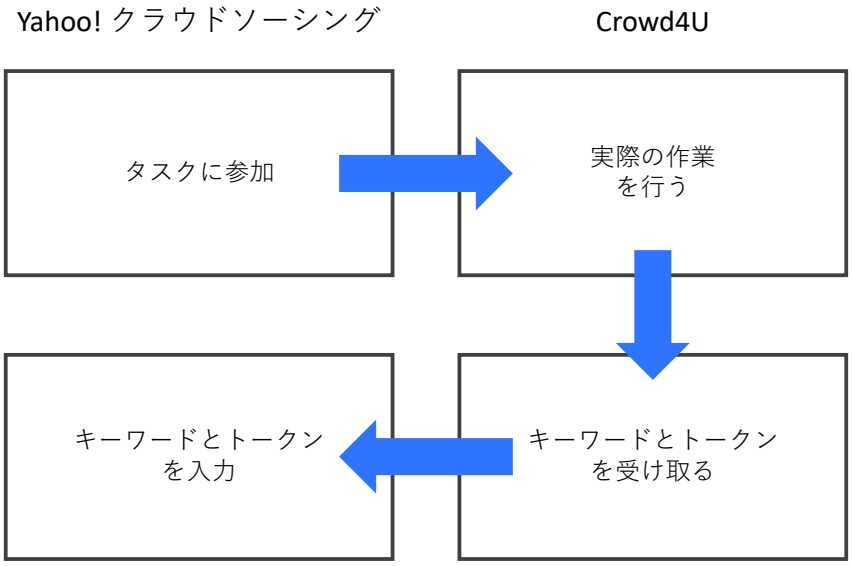


図 3.1　実験環境の概要

## 3.2　実験環境

### 3.2.1　実験環境の構成

実験環境の概略を 3.1 に示す．実験は Yahoo クラウドソーシング[*1]と Crowd4U[*2]を組み合わせて行った．ワーカの公募と報酬の支払いは Yahoo! クラウドソーシングを通じて行い，実際の作業ページは Crowd4U を用いて作成した．

実験に参加するワーカはまず，Yahoo! クラウドソーシング上のタスク一覧画面から，本研究にて作成した募集ページを選択し，作業に参加する．Yahoo クラウドソーシングにおける作業画面には，Crowd4U 上のタスク画面へのリンクとリンク先のページにて実際の作業を行う説明があり，リンク先のページへ移動して実際の作業を行う．作業が完了すると，Crowd4U 上の作業完了ページへと画面遷移する．作業完了ページにはキーワードとトークンが，Yahoo! クラウドソーシング上の画面にてこれらを入力するという説明とともに表示されている．最後に，Yahoo! クラウドソーシング上の作業画面にてキーワードとトークンを入力し，作業を完了させることで，報酬を受け取る．

Crowd4U 上での作業完了後に表示されるキーワードとトークンは最後まで作業に取り組み，報酬を受け取ったワーカを識別するためのものである．キーワードはすべてのワーカに共通な文字列であり，この文字列を得たワーカは作業を完了したとみなす．トークンはワーカごとに一意な文字列であり，個々のワーカを識別するためのものである．図 3.2 に Crowd4U での作業完了後に表示されるキーワードおよびトークンの表示画面の一例を示す．

[*1] https://crowdsourcing.yahoo.co.jp
[*2] https://crowd4u.org


図 3.2　Crowd4U での作業完了後に表示されるキーワードおよびトークンの表示画面

### 3.2.2　クラウドワーカの募集

Yahoo! クラウドソーシング上で報酬ありの作業として掲載することで参加者を公募した．作業に関する説明は日本語で記述されているため，実験参加者の多くは日本人であるか，日本語が理解できるようなワーカであると想定される．実験に最後まで参加した参加者には，回答の品質に関わらず 100 円相当の報酬を支払った．

### 3.2.3　各実験の設定の比較

図 3.3 に，本研究で取り組む各実験の全体としての目的を示す．全体の目的は，自己補正を現実のクラウドソーシングに適用した場合の，各自己補正タスクの正答率の改善（自己補正の短期的効果）と自己補正の繰り返しによるワーカ自身の正答率の改善（自己補正の長期的効果）が見られるかを検証することである．

表 3.1 に，各実験において設定が異なる点を示す．実験 1A では，自己補正における参考回答の有無が，自己補正の短期的効果および長期的効果にあたえる影響を検証する．実験では，実験に参加するワーカを 2 つのグループに分割し，片方のグループを自己補正における参考回答がありの条件に，もう一方のグループを参考回答が無しの条件に割り当てる．実験では鳥の画像分類課題を用いた．

実験 1B では，自己補正における参考回答の品質が，自己補正の短期的効果および長期的効果にあたえる影響を検証する．実験では，実験に参加するワーカを 2 つのグループに分割し，片方のグループを参考回答が常に正解の条件に，もう一方のグループを参考回答が選択肢からランダムに選んだ回答とする条件に割り当てる．実験では，絵画を提示してその作者を選択肢から選ぶ分類課題を用いた．

実験 2 では，自己補正における長期的効果が，学習に用いる課題と評価に用いる課題が異なる場合にも正答率の改善をもたらすか（学習の転移）を検証する．実験では，事前に平均正答率をクラウドソーシングにより測定したデータセットを複数組み合わせて，学習の転移がみられるかを明らかにする．比較する条件は，自己補正にて正解の参考回答を提示する場合と，自己補正を適用しない通常のタスクの場合を比較する．

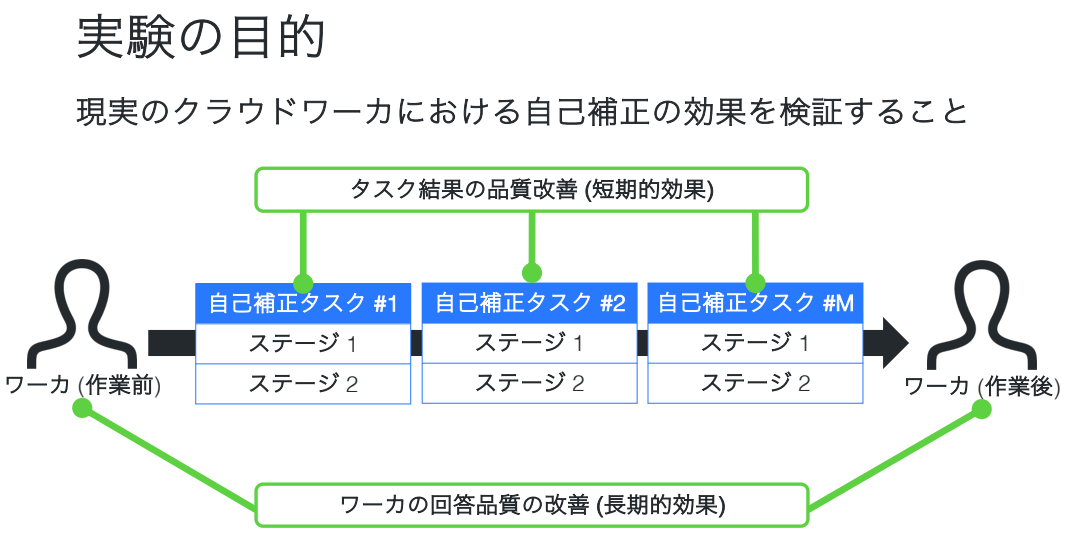


図 3.3　実験全体での目的

表 3.1　各実験の設定の比較

| | 実験 1A | 実験 1B | 実験 2 |
|---|---|---|---|
| 目的 | 自己補正の参考回答<br>の有無が重要か確かめる | 自己補正の参考回答<br>の品質が重要か確かめる | 自己補正による学習<br>の転移が見られるか<br>確かめる |
| 比較する条件 | 参考回答あり と<br>参考回答なし | 参考回答が正解 と<br>参考回答がランダム | 自己補正タスク と<br>通常のタスク |
| 学習タスク数 | 28 | 52 | 48 |
| タスク | 鳥の画像分類 | 絵画の分類 | 鳥の画像分類 |

# 第 4 章

# 実験 1（自己補正の短期的・長期的効果）

## 4.1　実験 1A（参考回答の有無の影響）

### 4.1.1　目的

クラウドソーシングにおいて，成果物の品質を改善することは重要な研究課題の 1 つである．Shah らが提案した自己補正は，作業に取り組むワーカに対して回答の機会を 2 度与えることでタスク結果の品質を改善する手法である．Shah らの論文では自己補正について，シミュレーションによる評価が行われたが，現実のクラウドワーカやタスクに対しても同様の傾向が見られるかは明らかでない．そこで本実験では，現実のクラウドワーカに自己補正を適用することで，現実のクラウドワーカのタスク結果が改善されるかを検証することを目的とし，画像分類課題を行った．

自己補正を適用したタスクにより現実のワーカが自身の回答を改善するならば，実際のクラウドソーシングにおいて，リクエスタが自己補正の導入を検討出来るようになるだろう．自己補正は，既存のタスク結果の改善手法である，多数決のようなタスク結果の集約や，優れたワーカを見つけて優先してタスクを割り当てるなどの手法とも組み合わせることが容易であることから，多くの場面で活用できると考えられる．

現実のクラウドソーシングにおいて自己補正を導入する場合，どの様な参考回答を提示するかを検討する必要がある．クラウドソーシングでは正解が未知の課題を扱うことが一般的であるため，正解を提示することは難しいため，すでに同様のタスクに回答したワーカの回答やその集約結果などを提示することになる．ただし，自己補正におけるタスク結果の品質改善において，2 段階のタスクデザインと参考回答の提示の両者が重要な要因であるかは自明ではない．そこで本実験では，参考回答の有無により，タスク結果の品質改善にどの様な影響をもたらすかを検証する．

自己補正によるタスク結果の改善は，2 段階目の回答の品質を改善することを目的とする手法である．しかし，ワーカが自己補正を繰り返す場合，以降の第一段階の回答の品質にも影響を与えると考えられる．そこで，ワーカが自己補正タスクを繰り返す実験により，ワーカ自身の回答精度に影響をもたらすかを検証する．

要約すると，この実験では自己補正タスクに取り組んだ現実のワーカのタスク結果について，次の 3 点を検討する．

1. 現実のクラウドソーシングタスクに自己補正を導入することで，ワーカから得られる回答の品質が改善されるか
2. 自己補正タスクによる回答品質の改善のために，参考回答として提示する回答の有無は重要な要因であるか

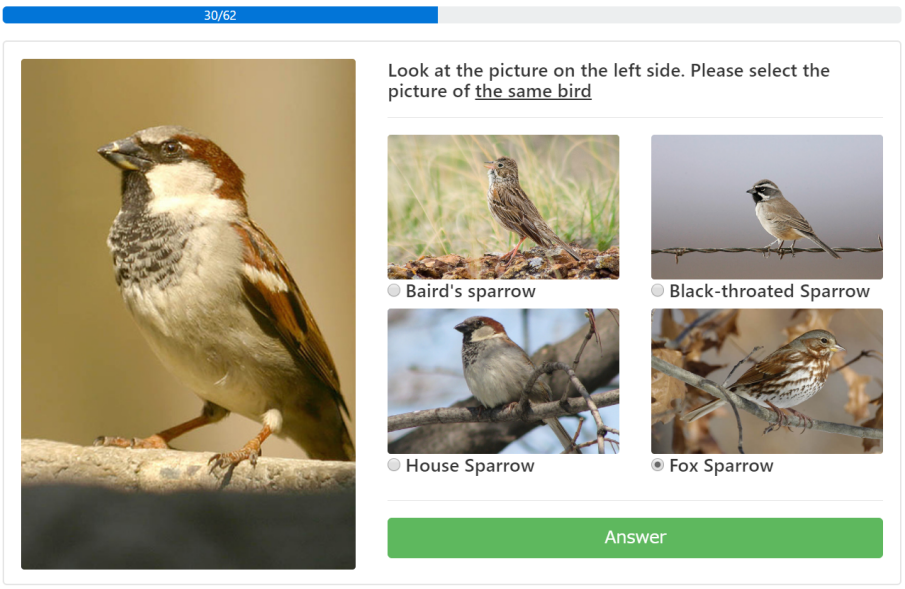


図 4.1　実験 1A で用いるテストタスクの一例

3. ワーカが自己補正タスクを繰り返すことにより，ワーカ自身の回答品質が改善されるか

### 4.1.2　実験方法

4 択の画像分類課題を用いた実験を行った．以下に実験の内容を詳述する．

**扱う課題**

実験参加者は選択式の画像分類タスクを行なった．選択肢は 4 種類で構成され，選択肢は全タスクを通して共通とした．タスクでは鳥類の画像のデータセットである Caltech-UCSD Birds 200[26] からを用いた．このデータセットには 200 種類の鳥について複数の画像が含まれている．データセットには鳥の種類毎に複数の画像が含まれているため，タスクの難易度を調節するために，容姿のよく似た種類の鳥を 4 種類選択した．提示される画像はワーカ間で共通であるが，出題する順番はワーカ毎に並び替えた．

**タスク**

実験では，提示された画像が与えられた 4 種類の選択肢のどの項目に該当するかを判断する画像分類課題を扱った．ワーカは，提示された画像が選択肢のどの項目に該当するかを推測し，その項目を選ぶ．実験では，自己補正を適用した自己補正タスクと，ワーカの評価のためのテストタスクを組み合わせて用いる．以下では，それぞれのタスクについて詳述する．

**■テストタスク**　テストタスクの例を図 4.1 に示す．テストタスクでは，ワーカは与えられた画像に対して単に分類作業を行う．選択肢の各項目の画像またはテキストをクリックすることで，回答とする項目を選ぶことが出来る．選択が済んだ後に，回答ボタンを押すことで，次のタスクへ遷移，または一連の作業が完了する．

**■自己補正タスク**　自己補正タスクの例を図 4.2 に示す．自己補正タスクにおいても，テストタスクと同様の画像分類課題を行う．自己補正タスクでは，選択肢を選ぶ機会が 2 回与えられる

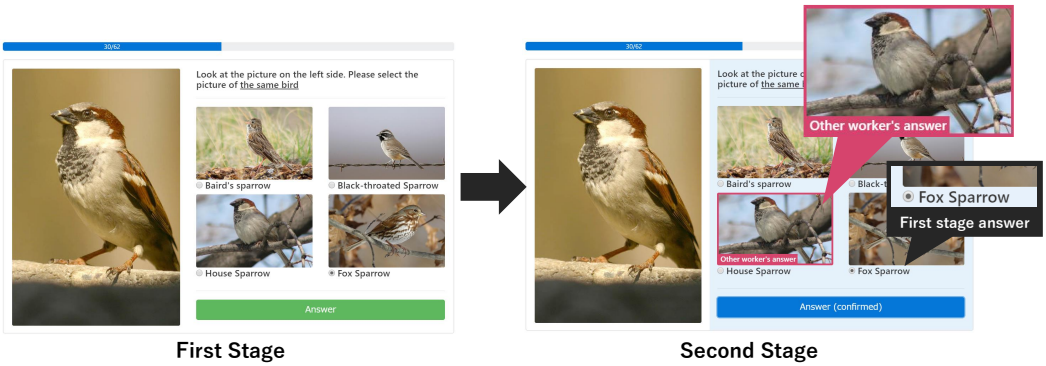


図 4.2　実験 1A で用いる自己補正タスクの一例

点が異なる．以降では各回答の機会についてステージ 1，ステージ 2 と呼ぶ．ステージ 1 において，選択肢からいずれかの項目を選択し，回答ボタンを押すことで，ステージ 2 の画面が表示される．ステージ 2 では，ステージ 1 でのワーカ自身の回答がチェックボックスに維持されているのに加え，参考回答である項目が赤枠でハイライトされる．ステージ 2 において，ワーカは自身の回答と参考回答を見た上で，最終的な回答を判断することが出来る．ステージ 2 で提示する参考回答には同じタスクに回答した別のワーカの回答を用いる．

### 比較する条件

実験では，自己補正のステージ 2 で提示する参考回答の重要性を明らかにするために，参考回答を提示する条件（以降では With reference 条件と呼ぶ）と参考回答を提示しない条件（以降では Without reference 条件と呼ぶ）についてタスクの正答率や反応時間などを比較する．以下に，それぞれの条件について詳述する．

■Without reference 条件　この条件では，自己補正のステージ 2 において，参考回答を提示しない．ワーカはステージ 2 において，ステージ 1 での自身の回答のみが確認できる．この条件では，ステージ 2 における参考回答のハイライトは行われない．

■With reference 条件　この条件では，自己補正のステージ 2 において，参考回答を提示する．ワーカはステージ 2 において，ステージ 1 での自身の回答に加えて，同じ質問に回答した他者の回答が確認できる．ステージ 2 での参考回答は赤い枠で示される．提示する参考回答には，Without reference 条件に参加したワーカについて，全タスクの正答率を算出し，上位 20% のワーカのタスク結果を用いる．正答率が上位 20% に該当したワーカのうち，ランダムに選ばれたワーカの回答が自己補正のステージ 2 にて提示される．この条件における参考回答はあくまで他者の回答として提示され，提示された参考回答の信頼性に関してワーカには事前に知ることが出来ないものとする．

### 実験デザイン

実験でワーカが取り組むタスクの構成を表 4.1 に示す．ワーカは一連の実験の通して 2 種類で構成される 5 つのフェーズのタスクに順番に回答する． 2 種類のフェーズの 1 つ目は，テストフェーズである．このフェーズではテストタスクが提示される． 2 つ目は，学習フェーズである．このフェーズでは自己補正タスクが提示される． 2 つのテスト時期の間に学習フェーズを割り当てることで，学習の効果を測定する．フェーズの構成は全てのワーカに対して共通であるが，出題するタスクや出題の順番はワーカごとにランダムに割り当てた．

表 4.1　実験の構成

| | フェーズ | タスクの種類 | タスク数 |
|---|---|---|---|
| 1 | Pre テスト | テスト | 12 |
| 2 | 学習 1 | 自己補正 | 28 |
| 3 | Mid テスト | テスト | 12 |
| 4 | 学習 2 | 自己補正 | 28 |
| 5 | Post テスト | テスト | 12 |

表 4.2　Pre テストの成績

| 条件 | フィルタ | N | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|---|
| Without reference | None | 98 | 0.826 | 0.148 | 0.25 | 0.75 | 0.833 | 0.917 | 1.0 |
| | Under 25% | 84 | 0.83 | 0.136 | 0.333 | 0.75 | 0.833 | 0.917 | 1.0 |
| With reference | None | 98 | 0.816 | 0.132 | 0.417 | 0.75 | 0.833 | 0.917 | 1.0 |
| | Under 25% | 86 | 0.824 | 0.134 | 0.417 | 0.75 | 0.833 | 0.917 | 1.0 |

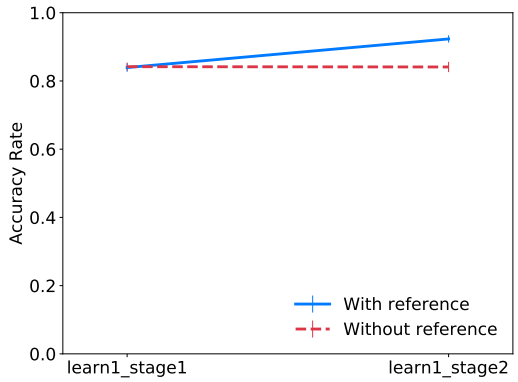


図 4.3　参考回答の各条件における，自己補正の各ステージの正答率（学習フェーズ 1 ）

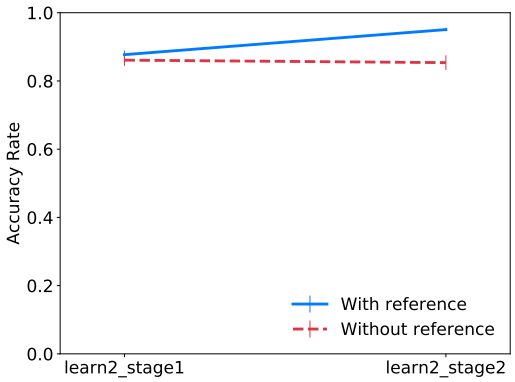


図 4.4　参考回答の各条件における，自己補正の各ステージの正答率（学習フェーズ 2）

### 4.1.3　実験結果

実験参加者 196 名からのデータが得られた．得られたデータのうち 26 名のデータを除外してデータの分析を行った．実験では，ワーカは複数のタスクに連続で取り組むため，途中から無作為に回答を選ぶようなワーカが見られた．そこで，mid テスト及び post テストの平均正答率が 25% を下回るワーカについては分析の対象から除外した．表 4.2 に条件毎の実験参加者数と除外した人数，pre テストの正答率の平均値を示す．

#### 自己補正の短期的効果

参考回答の条件毎の，学習フェーズ 1 における自己補正の各ステージの正答率を図 4.3 に示す．同様に，学習フェーズ 2 における自己補正の各ステージの正答率を図 4.4 に示す．図 4.3 および 4.4 の横軸は自己補正の各ステージで，縦軸は正答率を表している．

参考回答および学習フェーズ 1 での自己補正タスクのステージの違いによってタスクの正答

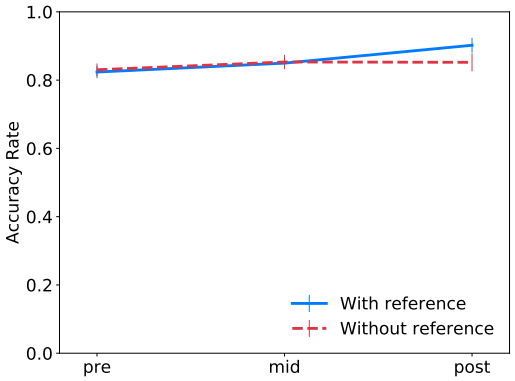


図 4.5　参考回答の各条件における，テスト時期ごとの正答率

率に差があるかを検証するために，独立変数を参考回答とステージ，従属変数をタスクの正答率とする混合計画の 2 要因の分散分析を行った．その結果，参考回答要因の主効果およびステージ要因の主効果，そして交互作用が有意であった ( $F(1, 168) = 6.578$, $p < .05$; $F(1, 168) = 33.74$, $p < .001$; $F(1, 168) = 34.31$, $p < .001$ )．まず，ステージ期要因の各水準における参考回答要因の単純主効果の検定を行ったところ，ステージ 1 水準においては単純主効果が認められなかったが，ステージ 2 水準では有意な単純主効果が認められた ( $F(1, 168) = .023$, n.s.; $F(1, 168) = 22.37$, .001 )．次に，参考回答要因の各水準におけるステージ要因の単純主効果の検定を行ったところ，With reference 水準では有意な単純主効果が認められた ( $F(1, 85) = 39.63$, $p < .001$ ) が Without reference 水準では単純主効果が認められなかった ( $F(1, 83) = 0.026$, n.s. )．

参考回答および学習フェーズ 2 での自己補正タスクのステージの違いによってタスクの正答率に差があるかを検証するために，独立変数を参考回答とステージ，従属変数をタスクの正答率とする混合計画の 2 要因の分散分析を行った．その結果，参考回答要因の主効果およびステージ要因の主効果，そして交互作用が有意であった ( $F(1, 168) = 10.9$, $p < .01$; $F(1, 168) = 33.19$, $p < .001$; $F(1, 168) = 47.86$, $p < .001$ )．まず，ステージ期要因の各水準における参考回答要因の単純主効果の検定を行ったところ，ステージ 1 水準においては単純主効果が認められなかったが，ステージ 2 水準では有意な単純主効果が認められた ( $F(1, 168) = .761$, n.s.; $F(1, 168) = 30.04$, .001 )．次に，参考回答要因の各水準におけるステージ要因の単純主効果の検定を行ったところ，With reference 水準では有意な単純主効果が認められた ( $F(1, 85) = 51.89$, $p < .001$ ) が Without reference 水準では単純主効果が認められなかった ( $F(1, 83) = 1.725$, n.s. )．

参考回答および各学習フェーズでの自己補正タスクのステージ 1 においてタスクの正答率に差があるかを検証するために，独立変数を参考回答と学習フェーズ，従属変数をタスクの正答率とする混合計画の 2 要因の分散分析を行った．その結果，学習フェーズ要因の主効果に有意差が認められ ( $F(1, 168) = 16.239$, $p < .001$ )，参考回答要因と交互作用には有意差が認められなかった ( $F(1, 168) = .175$, n.s.; $F(1, 168) = 1.731$, n.s )．

#### 自己補正の長期的効果

参考回答の条件毎の，各テスト時期の正答率を図 4.5 に示す．図 4.5 の横軸はテスト時期で，縦軸は正答率を表している．

参考回答およびテスト時期の違いによってタスクの平均正答率に差があるかを検証するため

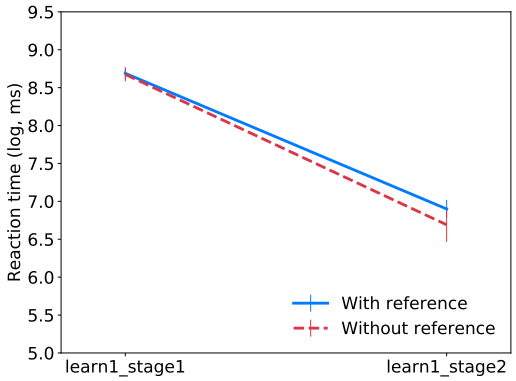


図 4.6　参考回答の各条件における，自己補正の各ステージの反応時間（学習フェーズ 1 ）

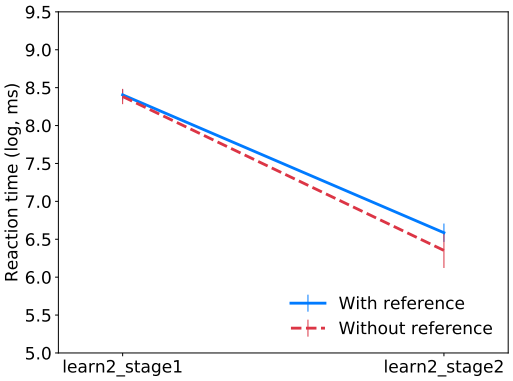


図 4.7　参考回答の各条件における，自己補正の各ステージの反応時間（学習フェーズ 2）

に，独立変数を参考回答とテスト時期，従属変数をタスクの正答率とする混合計画の 2 要因の分散分析を行った．その結果，テスト時期要因の主効果および交互作用が有意であった ( $F(2, 336) = 8.831$, $p < .001$; $F(2, 336) = 3.5$, $p < .05$; ) が，参考回答要因の主効果は有意ではなかった ( $F(1, 168) = 0.635$, n.s. )．まず，テスト時期要因の各水準における参考回答要因の単純主効果の検定を行ったところ，post テストにおいて有意な単純主効果が認められたが，pre テストおよび mid テストでは単純主効果が認められなかった ( pre: $F(1, 168) = .105$, n.s.; mid: $F(1, 168) = .027$, n.s.; post: $F(1, 168) = 4.475$, $p < .05$ )．次に，参考回答要因の各水準におけるテスト時期要因の単純主効果の検定を行ったところ，With reference 水準では有意な単純主効果が認められた ( $F(2, 170) = 11.82$, $p < .001$ ) が Without reference 水準では単純主効果が認められなかった ( $F(2, 166) = 1.08$, n.s. )．With reference 水準における各テスト時期に対してボンフェローニの方法による多重比較を行ったところ，post 水準と pre 水準の間，および post 水準と mid 水準の間に有意差が認められ，mid 水準と pre 水準の間には有意差が認められなかった．

反応時間（ステージ要因）

参考回答の条件毎の，学習フェーズ 1 における自己補正の各ステージの反応時間を図 4.6 に示す．同様に，学習フェーズ 2 における自己補正の各ステージの反応時間を図 4.7 に示す．図 4.6 および 4.7 の横軸は自己補正の各ステージで，縦軸は反応時間を表している．

参考回答および学習フェーズ 1 での自己補正タスクのステージの違いによってタスク回答での反応時間に差があるかを検証するために，独立変数を参考回答とステージ，従属変数をタスクの反応時間とする混合計画の 2 要因の分散分析を行った．その結果，参考回答要因の主効果およびステージ要因の主効果，そして交互作用が有意であった ( $F(1, 168) = 5.17$, $p < .05$; $F(1, 168) = 5814.9$, $p < .001$; $F(1, 168) = 15.3$, $p < .001$ )．まず，ステージ期要因の各水準における参考回答要因の単純主効果の検定を行ったところ，ステージ 1 水準においては単純主効果が認められなかったが，ステージ 2 水準では有意な単純主効果が認められた ( $F(1, 168) = .114$, n.s.; $F(1, 168) = 10.71$, .001 )．次に，参考回答要因の各水準におけるステージ要因の単純主効果の検定を行ったところ，With reference 水準と Without reference 水準のそれぞれに有意な単純主効果が認められた ( $F(1, 85) = 3604$, $p < .001$, $F(1, 85) = 2502$, $p < .001$ )．

参考回答および学習フェーズ 2 での自己補正タスクのステージの違いによってタスク回答での反応時間に差があるかを検証するために，独立変数を参考回答とステージ，従属変数をタス

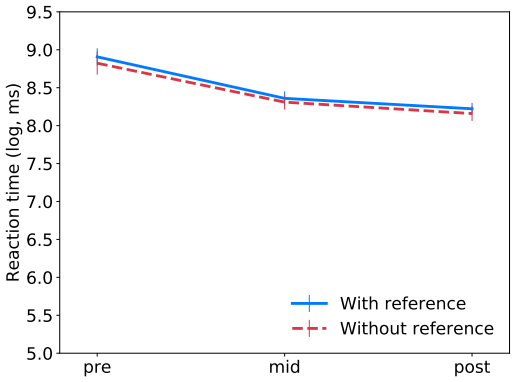


図 4.8　参考回答の各条件における，テスト時期ごとの反応時間

クの反応時間とする混合計画の 2 要因の分散分析を行った．その結果，参考回答要因の主効果およびステージ要因の主効果，そして交互作用が有意であった ( $F(1, 168) = 6.581$, $p < .05$; $F(1, 168) = 6227.94$, $p < .001$; $F(1, 168) = 18.76$, $p < .001$ )．まず，ステージ期要因の各水準における参考回答要因の単純主効果の検定を行ったところ，ステージ 1 水準においては単純主効果が認められなかったが，ステージ 2 水準では有意な単純主効果が認められた ( $F(1, 168) = .246$, n.s.; $F(1, 168) = 13.17$, .001 )．次に，参考回答要因の各水準におけるステージ要因の単純主効果の検定を行ったところ，With reference 水準と Without reference 水準のそれぞれに有意な単純主効果が認められた ( $F(1, 85) = 3285$, $p < .001$, $F(1, 85) = 2993$, $p < .001$ )．

参考回答および各学習フェーズでの自己補正タスクのステージ 1 における反応時間に差があるかを検証するために，独立変数を参考回答と学習フェーズ，従属変数をタスク回答の反応時間とする混合計画の 2 要因の分散分析を行った．その結果，学習フェーズ要因の主効果に有意差が認められ ( $F(1, 168) = 435.543$, $p < .001$ ) ，参考回答要因と交互作用には有意差が認められなかった ( $F(1, 168) = .192$, n.s.; $F(1, 168) = .068$, n.s )．

#### 反応時間（テスト時期要因）

参考回答の条件毎の，各テスト時期の反応時間を図 4.8 に示す．図 4.8 の横軸はテスト時期で，縦軸は反応時間を表している．

参考回答およびテスト時期の違いによってタスク回答の平均反応時間に差があるかを検証するために，独立変数を参考回答とテスト時期，従属変数をタスク回答の反応時間とする混合計画の 2 要因の分散分析を行った．その結果，テスト時期要因の主効果が有意であった ( $F(2, 336) = 614.783$, $p < .001$ ) が，参考回答要因の主効果および交互作用は有意ではなかった ( $F(1, 168) = 2.255$, n.s.; $F(1, 168) = 0.356$, n.s. )．各テスト時期に対してボンフェローニの方法による多重比較を行ったところ，post 水準と pre 水準の間および post 水準と mid 水準の間，mid 水準と pre 水準の間のすべてに有意差が認められた．

#### ワーカの成長度合い

ワーカの成長度合いを，post テストの正答率から pre テストの正答率を引いた値として考える．参考回答要因の各条件における，ワーカの成長度合いの分布を図 4.9 に示す．横軸は成長度合いを，縦軸は該当するワーカの割り合いを示す．

参考回答要因の各水準における成長度合いの分布に差があるかを確認するために，マンホイットニーの U 検定を行ったところ有意差が認められた ($p < .05$)．

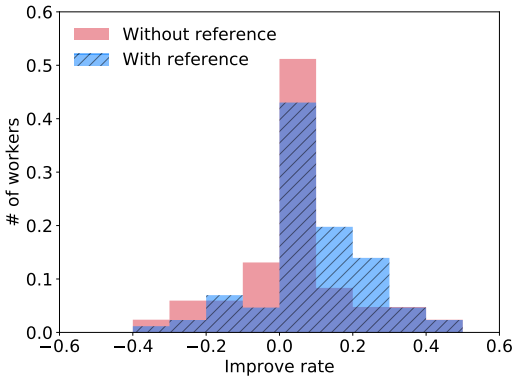


図 4.9　参考回答の条件毎の，ワーカの成長度合いの分布

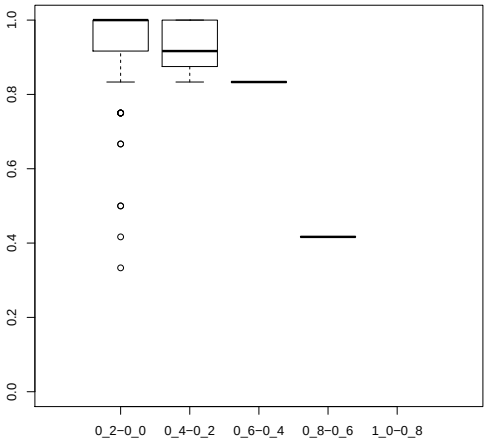


図 4.10　With reference 条件における回答変更率と post テストの正答率

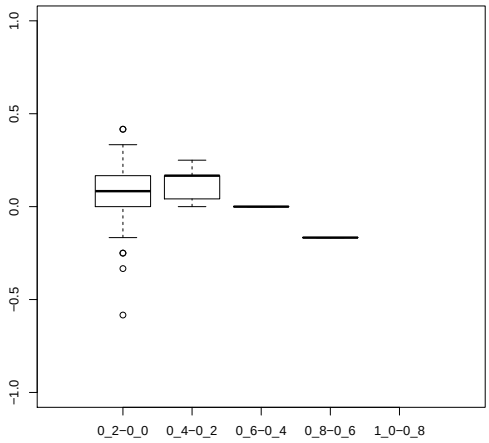


図 4.11　With reference 条件における回答変更率とワーカの成長度合い

**回答変更率**

自己補正の各ステージの回答の変更に着目する．ステージ 1 の回答が参考回答と異なる場合を式 4.1 に定義する．自己補正タスクの各タスク結果について，式 4.1 が真となる場合のうち，ステージ 1 とステージ 2 がどれだけ異なるかという割り合いを 4.2 に定義する．

$$stg1\_noteq\_RA = stage1 \neq reference\_answer \tag{4.1}$$

$$\text{回答変更率} = \frac{count(stg1\_noteq\_RA \&\& stage2 \neq stage1)}{count(stg1\_noteq\_RA)} \tag{4.2}$$

With reference 条件における回答変更率と post テストの正答率の関係を図 4.1.3 に示す．回答変更率と成長度合いの関係を 4.1.3 に示す．横軸は回答変更率を表しており，0.2 毎にグループ化している．

回答変更率の各段階により post テストの正答率に差があるかを検証するために，独立変数を回答変更率の各段階，従属変数をタスクの正答率とする一要因の分散分析を行った．その結果，主効果の有意差が認められた ($F(2, 82) = 4.288$, $p < .01$) ．多重比較を行ったところ，0.8-0.6 と 0.2-0.0，0.8-0.6 と 0.4-0.2 の間に有意差が認められた．

回答変更率の各段階により成長度合いに差があるかを検証するために，独立変数を回答変更率の各段階，従属変数を成長度合いとする一要因の分散分析を行った．その結果，主効果の有意差が認められなかった ($F(2, 82) = .924$, n.s.) ．

### 4.1.4 考察

#### 自己補正の短期的効果

学習フェーズ 1 での結果について考察する．自己補正のステージ要因の各水準における参考回答要因の単純主効果の検定では，ステージ 1 では有意差が認められず，ステージ 2 では有意差が認められた．そして，参考回答要因の各水準における，自己補正のステージ要因の単純主効果の検定では，With reference 条件では有意差が認められたが，Without reference 条件では有意差が認められなかった．このことから，With reference 水準で用いた様な性質を持つ参考回答を提示することで，自己補正のステージ 2 での正答率が改善されたと言える．With reference 水準における各ステージ水準での正答率は，Shah らの主張であるタスクに自己補正を導入することによる正答率の改善を支持するものである．

自己補正タスクによる正答率の改善は，Without reference 条件では見られず，With reference 条件でのみ見られたことから，参考回答の有無は重要な要因の 1 つであると考えられ，より信頼性の高い参考回答を提示することが求められると言える．ただし，一部のワーカは自己補正のステージ 2 において，常に参考回答として提示された回答へと変更するような振る舞いをすると考えられる．加えて，ワーカが参考回答を活用できるかどうかは，タスク自体の難易度や，ワーカがタスクで問われている内容を理解しているかなどに依存すると考えられる．

学習フェーズ 2 においても，学習フェーズ 1 と同様の傾向が見られた．学習フェーズ 2 における傾向の考察については，学習フェーズ 1 と同様の説明が可能であると考えられるため，説明は省略する．

学習フェーズ 1 および学習フェーズ 2 での自己補正タスクのステージ 1 の正答率に主効果が認められた．しかし，参考回答要因の主効果および交互作用の有意差が認められなかった．このことから，学習フェーズ 2 の自己補正タスクのステージ 1 での正答率の改善はタスクへの回答の繰り返しによるものであると考えられる．

#### 自己補正の長期的効果

テスト時期要因の各水準における参考回答要因の単純主効果の検定では，pre 水準および mid 水準には有意差が認められず，post 水準では有意差が認められた．そして，参考回答要因の各水準におけるテスト時期要因の単純主効果の検定では，With reference 水準では有意差が認められ，Without reference 水準では有意差が認められなかった．つまり，With reference 水準で用いた様な性質を持つ参考回答を提示することで，pre 水準の正答率を比べて post 水準の成績が上回ったと言える．このことから，ワーカが信頼できる参考回答を提示する自己補正タスクに連続で取り組むことで，ワーカ自身の正答率が改善することが示唆された．視覚的な作業を行う能力は，繰り返すことにより，すなわち知覚的な学習によりその速度や精度が改善されることが知られている [1]．知覚学習は意図せず無自覚に生じるものであり [2]，複数の研究が，視覚的な分類課題における知覚学習を報告している [3]．このことから，

自己補正の長期的効果は無自覚な知覚学習として説明できると考えられる．

With reference 水準における各テスト時期の成績の多重比較では，mid 水準と pre 水準の間には有意差が認められず，post 水準と mid 水準の間および post 水準と pre 水準の間に有意差が認められた．このことから，ワーカの学習にはある程度の自己補正タスクの繰り返しが必要であると考えれる．ただし，この傾向は今回の実験の設定の範囲内で主張できることであり，自己補正を繰り返す回数やタスクで扱う課題などによって成長の度合いが左右されることが予想される．

#### 反応時間（ステージ要因）

学習フェーズ 1 における，自己補正のステージ要因の各水準における参考回答要因の単純主効果の検定では，ステージ 1 では有意差が認められず，ステージ 2 では有意差が認められた．そして参考回答要因の各水準におけるステージ要因の単純主効果の検定では，With reference 水準と Without reference 水準の両者に有意差が認められた．このことから，自己補正タスクで提示する参考回答の有無にかかわらず，ステージ 2 の反応時間はステージ 1 よりも短くなることが示唆された．更に，ステージ 2 における反応時間は Without reference 水準と比較して With reference 水準のほうが有意に長いことが示唆された．ワーカは参考回答を提示された際に，自身の回答と提示された回答のどちらがより正しいかを判断するため，参考回答を提示する条件のほうが反応時間が長くなったと考えられる．

学習フェーズ 2 においても，学習フェーズ 1 と同様の傾向が見られた．学習フェーズ 1 と同様の議論ができると考えられるため，説明は省略する．

学習フェーズ 1 および学習フェーズ 2 での自己補正タスクのステージ 1 の反応時間率に主効果が認められた．しかし，参考回答要因の主効果および交互作用の有意差が認められなかった．このことから，学習フェーズ 2 の自己補正タスクのステージ 1 での反応時間の短縮はタスクへの回答の繰り返しによるものであると考えられる．

#### 反応時間（テスト時期要因）

テスト時期要因の主効果が認められた．テスト時期要因の各水準の多重比較では，全ての水準間に有意差が認められた．ただし，参考回答要因の主効果および交互作用についての有意差は認められなかった．このことから，テスト時期要因の反応時間の短縮は，作業の繰り返しによるものであると考えられる．

#### ワーカの成長度合いの分布

テスト時期における post の成績から pre の成績を引いた値をワーカの成長度合いと考える．各ワーカの成長度合いについてのヒストグラムを図 4.9 に示す．参考回答が With reference の条件では，成長度合いが 0.2 から 0.4 に相当するワーカの数が，Without reference の条件よりも多いことが分かる．このことから，With reference の参考回答を提示したことにより，一部のワーカについては回答品質の改善に繋がったと考えられる．

#### 正答率が改善したワーカの分析

条件毎の成長度合いの分布に有意差が認められた．With reference 条件から，一部のワーカの正答率が改善したことが分かる．

どの様なワーカに成長が見られたかを調べるために，回答変更率とポスト成績の分析をしたところ，いくつかの段階の間に有意差が見られた．このことから，成長するようなワーカは，全ての回答を変更したり，全く変更するような行動をしていないことが分かる．今回の実験で

は，ある範囲の変更率のワーカの成績が高かったが，これは取り扱う課題などによって異なると考えられる．

成長度合いと回答変更率の分析をしたところ，有意差は見られなかった．しかし，横軸の値のピークが post テストの成績の場合と成長度合いの場合で異なることから，もともと正答率が高かったワーカと成長により正答率が改善したワーカがいることが予想される．

## 4.2 実験 1B（参考回答の品質の影響）

### 4.2.1 目的

実験 1A では，鳥の画像分類タスクに対して自己補正を適用する実験を行った．その結果，参考回答を提示することでタスク結果の改善が見られた．さらに，自己補正を繰り返すことで，ワーカ自身の回答品質も改善されることが示唆された．しかし，参考回答要因の条件に関わらず，pre テストの平均正答率が高くなるような課題であったため，ワーカ自身の正答率の改善の幅が小さかった．そこで，この実験では，タスクの平均正答率が実験 1A よりも低くなるような課題を扱い，画像分類課題を行った．

実験 1A の結果から，自己補正において参考回答はタスク結果の改善およびワーカの学習において重要な要素であることが示唆された．しかし，参考回答の品質が，タスク結果の改善およびワーカの学習に与える影響は不明である．そこでこの実験では，参考回答として常に正答を提示する Correct 水準と常にランダムな回答を提示する Random 水準を比較することで，参考回答の品質がもたらす影響を明らかにする．

加えて，実験 1A におけるワーカの学習は pre テスト及び mid テストでは見られず，mid テストおよび post テストで見られたことから，ある程度の自己補正の繰り返しが必要であると考えられる．そのためこの実験では，各学習フェーズで割り当てるタスク数を増やすことでワーカの学習の度合いがどの様に変化するかを明らかにする．

### 4.2.2 実験方法

以下に実験内容について詳述する．

#### 実験環境

全作業を完了したワーカであるかを識別するために用いるキーワードとトークンについて，すべてのワーカに共通なキーワードについては新たに生成した文字列を使用した．

#### 扱う課題

この実験では，絵画を提示し，その絵画が選択肢の項目のどの画家の作品であるかを推定する課題を作成し，用いた．絵画の画像データについては wikiart.org [*1]にて収集した．

#### タスク

実験 1A と同様にテストタスクと自己補正タスクを組み合わせて実験を行う．図 4.12 に自己補正タスクのステージ 2 の一例を示す．

[*1] https://www.wikiart.org/

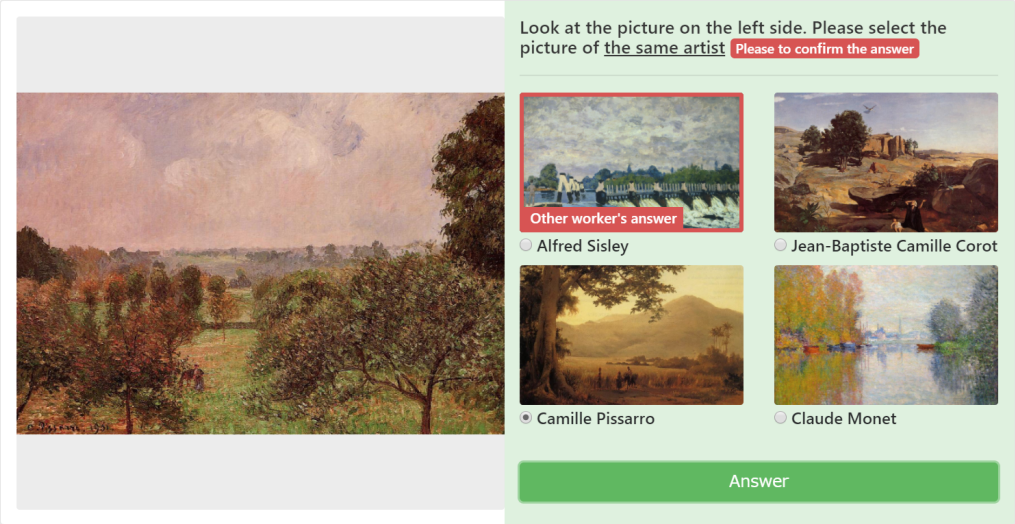


図 4.12　実験 1B で用いる自己補正タスクの一例（画像はステージ 2 の状態のみ）

表 4.3　実験の構成

| | フェーズ | タスクの種類 | タスク数 |
|---|---|---|---|
| 1 | Pre テスト | テスト | 12 |
| 2 | 学習 1 | 自己補正 | 52 + 2 回答既知タスク |
| 3 | Mid テスト | テスト | 12 |
| 4 | 学習 2 | 自己補正 | 52 + 2 回答既知タスク |
| 5 | Post テスト | テスト | 12 |

**比較する条件**

実験では，自己補正のステージ 2 で提示する参考回答の品質の重要性を明らかにするために，常にランダムな回答を見せる random 条件と，常に正答を提示する correct 条件について，タスク結果の正答率は反応時間を比較した．実験 1A と同様に，実験に参加するワーカはどちらの条件のグループに割り当てられるかについて告知しない．

**ワーカのフィルタリング**

実験 1 A では，ランダムな回答を行うワーカを分析の対象から除外するために，mid テストおよび post テストの成績に基づいてフィルタリングを行った．この実験では，ワーカをフィルタリングするためのタスクを導入し，それらのタスクに正答できたかどうかに基づいてワーカをフィルタリングする．フィルタリングのためのタスクは各学習フェーズに 2 タスクずつ追加した．タスクでは，質問として選択肢として提示されている画像が提示される．ワーカに対してはこれらのタスクが含まれていることは告知せず，フィルタリングにより分析の対象から除外される場合にも報酬を支払った．

**実験デザイン**

実験でワーカが取り組むタスクの構成を表 4.3 に示す．フェーズの構成自体は実験 1A と同様であるが，学習フェーズにおけるタスク数が異なる．

表 4.4　Pre テストの成績

| 条件 | フィルタ | N | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|---|
| random | None | 105 | 0.354 | 0.152 | 0.0833 | 0.25 | 0.333 | 0.417 | 1.00 |
| | Gold | 74 | 0.365 | 0.157 | 0.0833 | 0.25 | 0.333 | 0.417 | 1.00 |
| correct | None | 86 | 0.356 | 0.15 | 0.0833 | 0.25 | 0.333 | 0.417 | 0.75 |
| | Gold | 58 | 0.353 | 0.154 | 0.0833 | 0.25 | 0.333 | 0.417 | 0.75 |

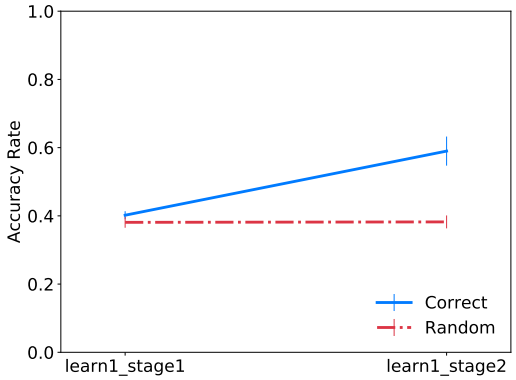


図 4.13　参考回答の各条件における，自己補正の各ステージの正答率 (学習フェーズ 1)

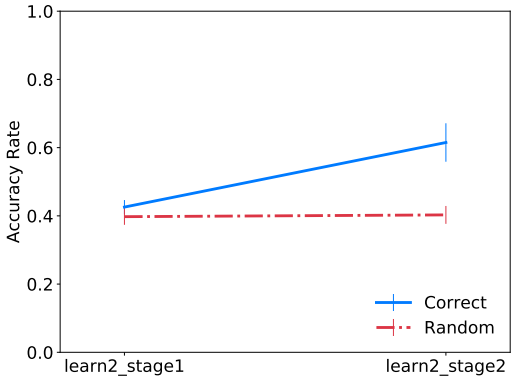


図 4.14　参考回答の各条件における，自己補正の各ステージの正答率 (学習フェーズ 2)

### 4.2.3　実験結果

実験参加者 191 名からのデータが得られた．得られたデータのうち，59 名のデータを除外してデータの分析を行った．除外したのは，学習フェーズ中に出題した，選択肢と同様の画像が出題されるタスクについて，正答出来なかったワーカである．表 4.4 に，条件毎の実験参加者数と分析から除外した人数，各セクションの正答率の平均値を示す．

#### 自己補正の短期的効果

参考回答の条件毎の，学習フェーズ 1 における自己補正の各ステージの正答率を図 4.13 に示す．同様に，学習フェーズ 2 における自己補正の各ステージの正答率を図 4.14 に示す．図 4.13 および 4.14 の横軸は自己補正の各ステージで，縦軸は正答率を表している．

参考回答および学習フェーズ 1 での自己補正タスクのステージの違いによってタスクの正答率に差があるかを検証するために，独立変数を参考回答とステージ，従属変数をタスクの正答率とする混合計画の 2 要因の分散分析を行った．その結果，参考回答要因の主効果およびステージ要因の主効果，そして交互作用が有意であった ( $F(1, 130) = 24.46$, $p < .001$; $F(1, 130) = 52.59$, $p < .001$; $F(1, 130) = 65.03$, $p < .001$ )．まず，ステージ期要因の各水準における参考回答要因の単純主効果の検定を行ったところ，ステージ 1 水準においては単純主効果が認められなかったが，ステージ 2 水準では有意な単純主効果が認められた ( $F(1, 130) = 1.008$, n.s.; $F(1, 130) = 47.75$, .001 )．次に，参考回答要因の各水準におけるステージ要因の単純主効果の検定を行ったところ，correct 水準では有意な単純主効果が認められた ( $F(1, 57) = 64.5$, $p < .001$ ) が random 水準では単純主効果が認められなかった ( $F(1, 73) = .02$, n.s. )．

参考回答および学習フェーズ 2 での自己補正タスクのステージの違いによってタスクの正答

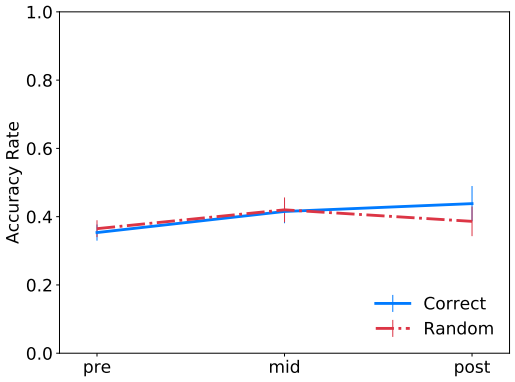


図 4.15　参考回答の各条件における，テスト時期ごとの正答率

率に差があるかを検証するために，独立変数を参考回答とステージ，従属変数をタスクの正答率とする混合計画の 2 要因の分散分析を行った．その結果，参考回答要因の主効果およびステージ要因の主効果，そして交互作用が有意であった ( $F(1, 130) = 17.6$, $p < .01$; $F(1, 130) = 57.44$, $p < .001$; $F(1, 130) = 64.71$, $p < .001$ )．まず，ステージ期要因の各水準における参考回答要因の単純主効果の検定を行ったところ，ステージ 1 水準においては単純主効果が認められなかったが，ステージ 2 水準では有意な単純主効果が認められた ( $F(1, 130) = 1.145$, n.s.; $F(1, 130) = 37.15$, .001 )．次に，参考回答要因の各水準におけるステージ要因の単純主効果の検定を行ったところ，correct 水準では有意な単純主効果が認められた ( $F(1, 57) = 57.9$, $p < .001$ ) が random 水準では単純主効果が認められなかった ( $F(1, 73) = .868$, n.s. )．

参考回答および各学習フェーズでの自己補正タスクのステージ 1 においてタスクの正答率に差があるかを検証するために，独立変数を参考回答と学習フェーズ，従属変数をタスクの正答率とする混合計画の 2 要因の分散分析を行った．その結果，学習フェーズ要因の主効果に有意差が認められ ( $F(1, 130) = 3.418$, $p < .01$ )，参考回答要因と交互作用には有意差が認められなかった ( $F(1, 130) = 1.349$, n.s.; $F(1, 130) = 0.112$, n.s )．

#### 自己補正の長期的効果

参考回答の条件毎の，各テスト時期の正答率を図 4.15 に示す．図 4.15 の横軸はテスト時期で，縦軸は正答率を表している．

参考回答およびテスト時期の違いによってタスクの平均正答率に差があるかを検証するために，独立変数を参考回答とテスト時期，従属変数をタスクの正答率とする混合計画の 2 要因の分散分析を行った．その結果，テスト時期要因の主効果が有意であった ( $F(2, 260) = 5.694$, $p < .01$ ) が，参考回答要因の主効果および交互作用は有意ではなかった ( $F(1, 130) = .226$, n.s.; $F(2, 260) = 1.738$, n.s. )．テスト時期要因に対してボンフェローニの方法による多重比較を行ったところ，pre 水準と mid 水準の間には有意差が認められたが ($p < 0.05$)，pre 水準および post 水準の間と mid 水準および post 水準の間には有意差が認められなかった．

#### 反応時間（ステージ要因）

参考回答の条件毎の，学習フェーズ 1 における自己補正の各ステージの反応時間を図 4.16 に示す．同様に，学習フェーズ 2 における自己補正の各ステージの反応時間を図 4.17 に示す．図 4.16 および 4.17 の横軸は自己補正の各ステージで，縦軸は反応時間を表している．

参考回答および学習フェーズ 1 での自己補正タスクのステージの違いによってタスク回答で

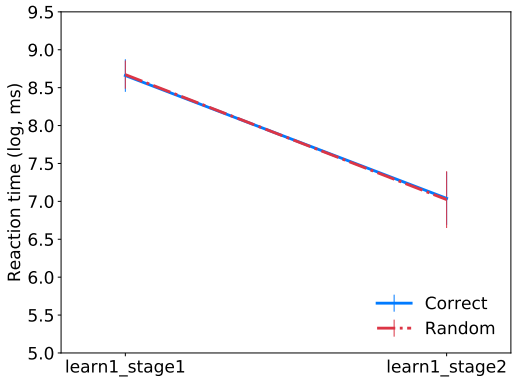


図 4.16　参考回答の各条件における，自己補正の各ステージの反応時間 (学習フェーズ 1)

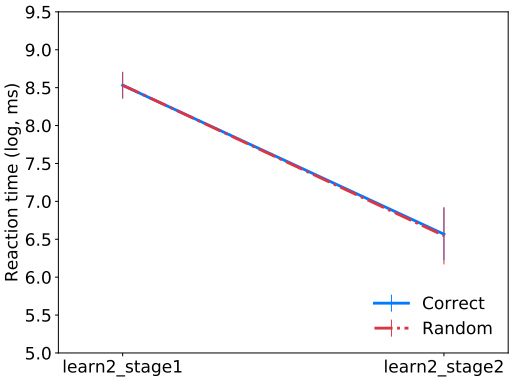


図 4.17　参考回答の各条件における，自己補正の各ステージの反応時間 (学習フェーズ 2)

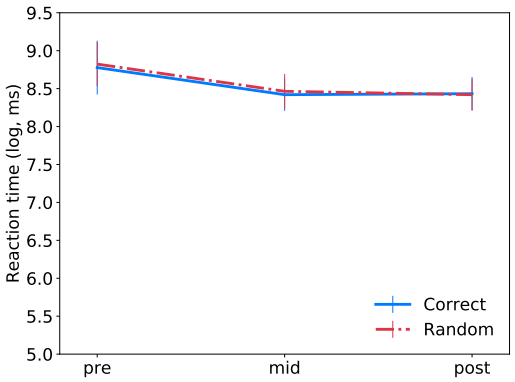


図 4.18　参考回答の各条件における，テスト時期ごとの反応時間

の反応時間に差があるかを検証するために，独立変数を参考回答とステージ，従属変数をタスクの反応時間とする混合計画の 2 要因の分散分析を行った．その結果，ステージ要因の主効果が有意であった ( $F(1, 130) = 2147.391$, $p < .001$ ) が，参考回答要因の主効果および交互作用には有意差が認められなかった ( $F(1, 130) = .001$, n.s.; $F(1, 130) = .159$, n.s. )．

参考回答および学習フェーズ 2 での自己補正タスクのステージの違いによってタスク回答での反応時間に差があるかを検証するために，独立変数を参考回答とステージ，従属変数をタスクの反応時間とする混合計画の 2 要因の分散分析を行った．その結果，ステージ要因の主効果が有意であった ( $F(1, 130) = 2449.221$, $p < .001$ ) が，参考回答要因の主効果および交互作用には有意差が認められなかった ( $F(1, 130) = .011$, n.s.; $F(1, 130) = .071$, n.s. )..

参考回答および各学習フェーズでの自己補正タスクのステージ 1 における反応時間に差があるかを検証するために，独立変数を参考回答と学習フェーズ，従属変数をタスク回答の反応時間とする混合計画の 2 要因の分散分析を行った．その結果，学習フェーズ要因の主効果に有意差が認められ ( $F(1, 130) = 107.111$, $p < .001$ )，参考回答要因と交互作用には有意差が認められなかった ( $F(1, 130) = .009$, n.s.; $F(1, 130) = .151$, n.s )．

反応時間（テスト時期要因）

参考回答の条件毎の，各テスト時期の反応時間を図 4.18 に示す．図 4.18 の横軸はテスト時期で，縦軸は反応時間を表している．

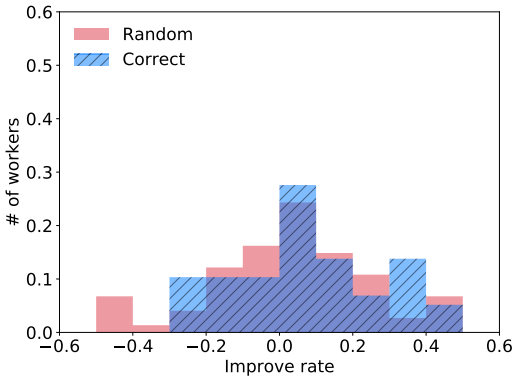


図 4.19　参考回答の条件毎の，ワーカの成長度合いの分布

参考回答およびテスト時期の違いによってタスク回答の平均反応時間に差があるかを検証するために，独立変数を参考回答とテスト時期，従属変数をタスク回答の反応時間とする混合計画の 2 要因の分散分析を行った．その結果，テスト時期要因の主効果が有意であった ( $F(2, 260) = 133.56$, $p < .001$ ) が，参考回答要因の主効果および交互作用は有意ではなかった ( $F(1, 130) = .098$, n.s.; $F(1, 130) = .83$, n.s. )．

各テスト時期に対してボンフェロー二の方法による多重比較を行ったところ，post 水準と pre 水準の間および mid 水準と pre 水準の間の有意差が認められたが，post 水準と mid 水準の間には有意差が認められなかった．

#### ワーカの成長度合い

ワーカの成長度合いを，post テストの正答率から pre テストの正答率を引いた値として考える．参考回答要因の各条件における，ワーカの成長度合いの分布を図 4.19 に示す．横軸は成長度合いを，縦軸は該当するワーカの割り合いを示す．

参考回答要因の各水準における成長度合いの分布に差があるかを確認するために，マンホイットニーの U 検定を行ったところ有意差は認められなかった．

#### 回答変更率

correct 条件における回答変更率と post テストの正答率の関係を図 4.20 に示す．correct 条件における回答変更率と成長度合いの関係を 4.21 に示す．同様に，random 条件における回答変更率と post テストの正答率の関係を図 4.22 に示す．random 条件における回答変更率と成長度合いの関係を 4.23 に示す．横軸は回答変更率を表しており，0.2 毎にグループ化している．

correct 条件における回答変更率の各段階により post テストの正答率に差があるかを検証するために，独立変数を回答変更率の各段階，従属変数をタスクの正答率とする一要因の分散分析を行った．その結果，主効果の有意差が認められた ($F(4, 53) = 7.78$, $p < .001$) ．多重比較を行ったところ，0.6-0.4 と 0.2-0.0，0.8-0.6 と 0.2-0.0，0.8-0.6 と 0.4-0.2，1.0-0.8 と 0.8-0.6，の間に有意差が認められた．

correct 条件における回答変更率の各段階により成長度合いに差があるかを検証するために，独立変数を回答変更率の各段階，従属変数を成長度合いとする一要因の分散分析を行った．その結果，主効果の有意差は認められなかった．

random 条件における回答変更率の各段階により post 時期の正答率に差があるかを検証す

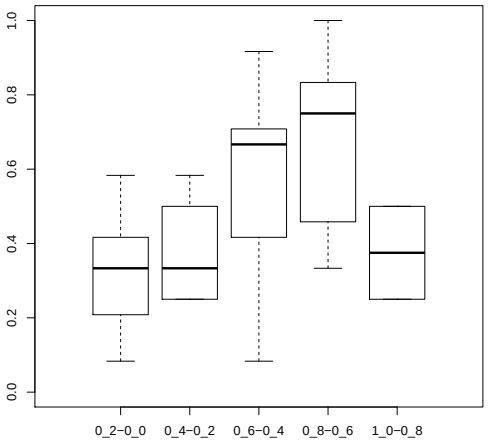


図 4.20　correct 条件における回答変更率と post テストの正答率

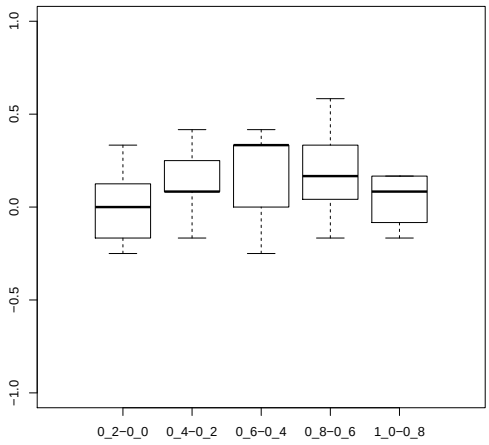


図 4.21　correct 条件における回答変更率と成長度合いの正答率

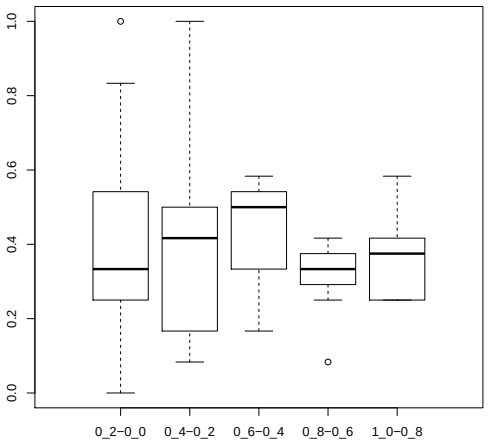


図 4.22　random 条件における回答変更率と post テストの正答率

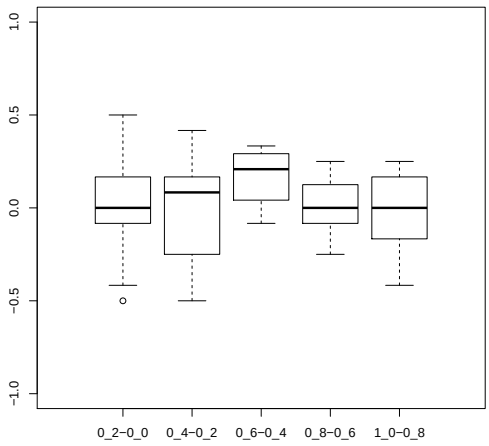


図 4.23　random 条件における回答変更率と成長度合いの正答率

るために，独立変数を回答変更率の各段階，従属変数を正答率とする一要因の分散分析を行った．その結果，主効果の有意差は認められなかった ($F(4, 69) = .31$, n.s.)．

random 条件における回答変更率の各段階により成長度合いに差があるかを検証するために，独立変数を回答変更率の各段階，従属変数を成長度合いとする一要因の分散分析を行った．その結果，主効果の有意差は認められなかった ($F(4, 69) = .518$, n.s.)．

### 4.2.4　考察

**自己補正の短期的効果**

学習フェーズ 1 および学習フェーズ 2 のそれぞれにおいて，自己補正のステージ要因の各水準における参考回答要因の単純主効果の検定では，ステージ 1 では有意差が認められず，ス

テージ 2 では有意差が認められた．そして，参考回答要因の各水準における，ステージ要因の単純主効果の検定では，correct 水準では有意差が認められたが，random 水準では有意差が認められなかった．correct 水準における結果は，Shah らの主張であるタスクに自己補正を導入することによる正答率の改善および，実験 1A の結果を支持するものである．このことから，参考回答の品質は自己補正がもたらす効果を左右する要因であることが示唆された．実験 1B では参考回答としてランダムおよび正答を，実験 1A では現実のワーカから得た回答を提示する実験を行ったが，参考回答の品質と自己補正によるタスク結果の改善の関係についてはこれらの結果から説明することは出来ない．

correct 水準においては，ステージ 2 の正答率が 100% に近づくことも予想されたが，ステージ 1 の正答率である 40% から 60% への改善に留まった．そして，random 水準に場合においてはステージ 2 の正答率がチャンスレベルである 25% に近づくことが予想されたが，ステージ 2 の正答率はステージ 1 と同程度であった．このことから，大半のワーカは参考回答を見た上で，自身の回答を変更するかを判断していると考えられる．実験 1A および実験 1B では，参考回答を単に他者の回答であるとして提示したが，例えば信頼できるワーカの回答であるなどと説明を付け加えることにより，参考回答に変更する割り合いが変化する可能性がある．

学習フェーズ 1 および学習フェーズ 2 での自己補正タスクのステージ 1 の正答率に主効果が認められた．しかし，参考回答要因の主効果および交互作用の有意差が認められなかった．このことから，学習フェーズ 2 の自己補正タスクのステージ 1 での正答率の改善はタスクへの回答の繰り返しによるものであると考えられる．

#### 自己補正の長期的効果

参考回答要因とテスト時期要因の分散分析では，テスト時期要因の主効果にのみ有意差が認められた．多重比較では，pre 水準および mid 水準の間にのみ有意差が認められた．このことから，この実験では実験 1A で述べた様なワーカの学習の効果は見られなかった．この原因の 1 つとしては，タスクの難易度を高く設定したことにより，ワーカが各選択毎の特徴を捉えたり，提示された参考回答の信頼性を判断するのが難しくなっていることが挙げられる．

自己補正タスクを割り当てる回数をより増やすことにより，学習の機会を増やすことで，交互作用が認められる可能性もあるが，一方で作業を中断したり離脱するワーカの存在が懸念される．

#### 反応時間（ステージ要因）

学習フェーズ 1 および学習フェーズ 2 のそれぞれの分析において，ステージ要因の主効果のみに有意差が認められた．このことから，ステージ 2 での反応時間はステージ 1 の反応時間よりも短いことが示唆され，これは実験 1A の結果を支持するものである．一方で，交互作用が認められなかったことから，参考回答の品質による反応時間の変化は見られなかった．

学習フェーズ 1 および学習フェーズ 2 での自己補正タスクのステージ 1 の反応時間に主効果が認められた．しかし，参考回答要因の主効果および交互作用の有意差が認められなかった．このことから，学習フェーズ 2 の自己補正タスクのステージ 1 での反応時間の短縮はタスクへの回答の繰り返しによるものであると考えられる．

#### 反応時間（テスト時期要因）

テスト時期要因の反応時間要因の分散分析では，テスト時期要因の主効果のみが認められた．テスト時期要因の多重比較では，post 時期と pre 時期の間および pre 時期と mid 時期の間に有意差が認められたが，post 時期および mid 時期の間には有意差は認められなかった．

これらのテスト時期要因の反応時間の短縮は，作業の繰り返しによるものであると考えられる．

正答率が改善したワーカの分析

実験 1.1 のヒストグラムには有意差があった．一方で，実験 1.2 のヒストグラムには有意差がなかった．どんなワーカの post の成績が高いかを調べるために，回答変更率と post 成績の関係を調べた．その結果，0.6-0.4 と 0.2-0.0，0.8-0.6 と 0.2-0.0，0.8-0.6 と 0.4-0.2，1.0-0.8 と 0.8-0.6 に有意差が認められた．このことから，post の成績が高いワーカの回答変更率は著しく高くなく，一方で著しく低くならないという性質を持つことが示唆された．この結果は実験 1.1 での結果を支持するものである．

更に，post-pre と回答変更率の関係を調べたところ，有意差は認められなかった．ピークとなる値が異なることから，もともと成績が高いワーカと作業と通して正答率が改善したワーカの存在が示唆された．

同様の random 条件の post 成績と回答変更率の関係，および post-pre と回答変更率の関係を調べたところ，有意差は認められなかった．しかし，分布のピークのずれが With reference と類似していることがわかる．このことから，ワーカの集合には正答率が改善する可能性を持つワーカが潜在的に含まれており，その改善の度合いが参考回答要因の条件などによって変化した可能性が考えられる．

# 第 5 章

# 実験 2（自己補正による学習の転移）

## 5.1　実験 2

### 5.1.1　目的

実験 1A および実験 1B では，自己補正の短期的効果と長期的効果が，異なるデータセットによる 2 つの実験で見られた．これらの実験では，学習フェーズのタスクとテストフェーズのタスクが同じ選択肢で構成されていた．それに対して実験 2 では，学習フェーズのタスクとテストフェーズのタスクで異なる選択肢で構成されるタスクを割り当て，自己補正の長期の効果が，別の課題での正答率を改善するかを明らかにする．

### 5.1.2　実験方法

#### タスク

実験 1A と同様にテストタスクと自己補正タスクを組み合わせて実験を行う．図 5.1.2 に自己補正タスクのステージ 2 の一例を示す．

#### データセット

異なる選択肢で構成されるタスクを組み合わせて用いるにあたり，タスクの候補を複数作成し，Amazon Mechanical Turk[*1] を利用して実際に一部のタスクの作業を依頼した．集められたタスク結果から，各データセットの正答率の傾向を求め，傾向が類似しているデータセットのペアを選びだした．データセット 1 とデータセット 2 は，平均正答率が 50% 付近となるようなデータセットである．データセット 3 とデータセット 4 は平均正答率が 90% 付近となるようなデータセットである．実験では，データセット 1 と 2 をペアに，データセット 3 と 4 をペアにして，学習タスクとテストたすくにそれぞれを適用することで，合計 4 種類の課題を作成した．

#### 比較する条件

実験では，自己補正によりワーカの学習が別の課題でも有効であるかを明らかにするために，自己補正のステージ 2 で正答を提示する correct 条件と，自己補正ではなくテストタスクと同等の NSC 条件について，タスク結果の正答率を比較した．

[*1] https://www.mturk.com/


図 5.1　実験 2 で用いる自己補正タスクの一例（画面はステージ 2 の状態）

表 5.1　実験の構成

| | フェーズ | タスクの種類 | タスク数 |
|---|---|---|---|
| 1 | Pre テスト | テスト | 24 |
| 2 | 学習 1 | 自己補正 | 48 + 4 回答既知タスク |
| 3 | Post テスト | テスト | 24 |

**実験デザイン**

実験でワーカが取り組むタスクの構成を表 5.1 に示す．フェーズの要素自体は実験 1A と同様であるが，実験 1A では学習フェーズは 2 回であるが，この実験では 1 回である．更にテストフェーズで割り当てるタスク数が実験 1A とは異なる．

**ワーカのフィルタ**

実験 1-2 と同様の方法を用いた．learning フェーズに，選択肢に表示される画像が出題されるタスクを 4 回提示し，それら全てに回答できたワーカを分析の対象とした．

### 5.1.3　実験結果

**テスト要因の分析**

各データセットにおける参考回答の条件毎の，各テスト時期の正答率を図 5.2 に示す．これらの図の横軸はテスト時期を，縦軸は正答率を表している．

データセット 1 における，参考回答およびテスト時期によってタスクの正答率に差があるかを検証するために，独立変数を参考回答とテスト時期，従属変数をタスクの正答率とする混合計画の 2 要因の分散分析を行った．その結果，参考回答要因の主効果およびテスト時期要因の主効果，そして交互作用のすべてに有意差が認められなかった ( $F(1, 63) = .214$, n.s.; $F(1, 63) = .162$, n.s.; $F(1, 63) = .049$, n.s. ).

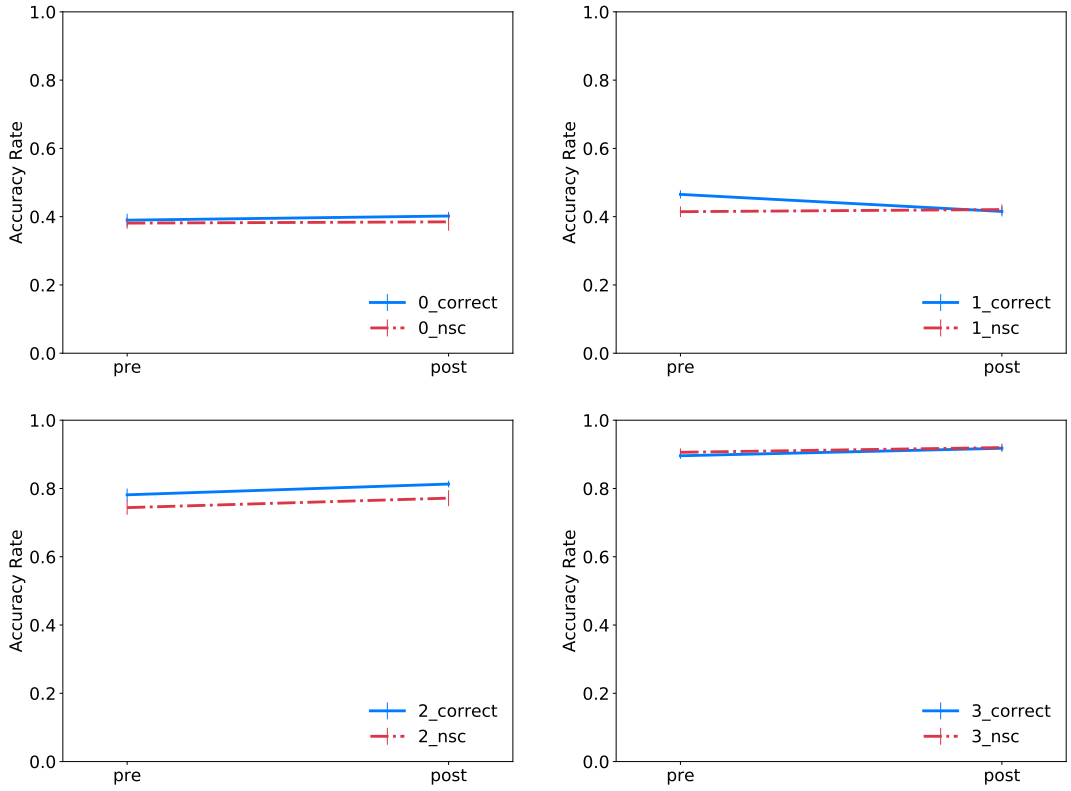


図 5.2　各データセットにおける，条件毎の各テスト時期の正答率

データセット 2 における，参考回答およびテスト時期によってタスクの正答率に差があるかを検証するために，独立変数を参考回答とテスト時期，従属変数をタスクの正答率とする混合計画の 2 要因の分散分析を行った．その結果，参考回答要因の主効果およびテスト時期要因の主効果，そして交互作用のすべてに有意差が認められなかった ( $F(1, 68) = .947$, n.s.; $F(1, 68) = 1.082$, n.s.; $F(1, 68) = 2.630$, n.s. ).

データセット 3 における，参考回答およびテスト時期によってタスクの正答率に差があるかを検証するために，独立変数を参考回答とテスト時期，従属変数をタスクの正答率とする混合計画の 2 要因の分散分析を行った．その結果，テスト時期要因の主効果が有意であった ( $F(1, 91) = 4.838$, $p < .05$ ) ，が参考回答要因の主効果および交互作用には有意差が認められなかった ( $F(1, 91) = 2.571$, n.s.; $F(1, 91) = .019$, n.s ).

データセット 4 における，参考回答およびテスト時期によってタスクの正答率に差があるかを検証するために，独立変数を参考回答とテスト時期，従属変数をタスクの正答率とする混合計画の 2 要因の分散分析を行った．その結果，参考回答要因の主効果およびテスト時期要因の主効果，そして交互作用のすべてに有意差が認められなかった ( $F(1, 77) = .09$, n.s.; $F(1, 77) = 3.327$, n.s.; $F(1, 77) = .165$, n.s. ).

#### 回答変更率の分析

各データセットの correct 水準における，post 時期の正答率と回答変更率の関係をを図に示す（データセット 1: 図 5.3, データセット 2: 図 5.4, データセット 3: 図 5.5, データセット 4: 図 5.6 ）．これらの図の横軸は回答変更率の各段階を，縦軸は正答率を表している．

それぞれのデータセットでの結果において，回答変更率により post 時期の成績に差がある

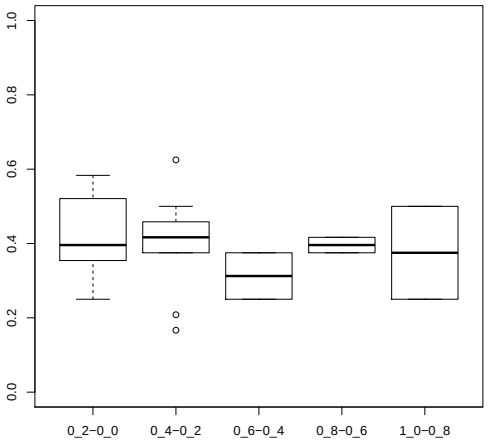


図 5.3　データセット 1 の correct 条件における回答変更率と post テストの正答率

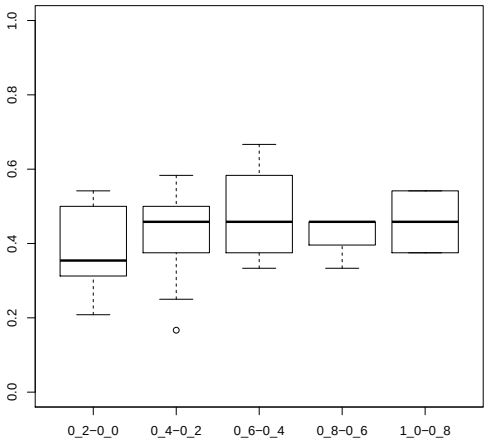


図 5.4　データセット 2 の correct 条件における回答変更率と post テストの正答率

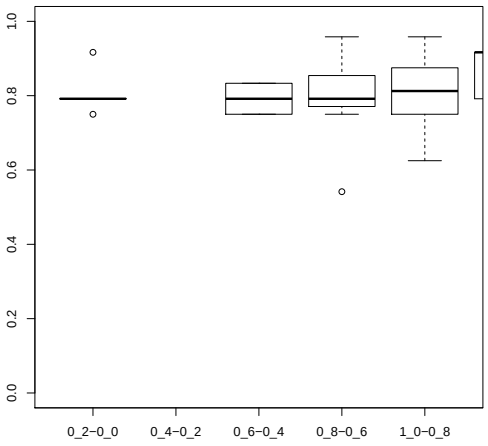


図 5.5　データセット 3 の correct 条件における回答変更率と post テストの正答率

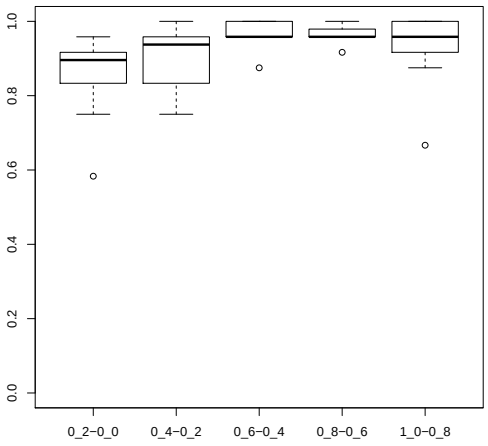


図 5.6　データセット 4 の correct 条件における回答変更率と post テストの正答率

かを検証するために，独立変数を rate，従属変数を正答率とする一要因の分散分析を行った．その結果，データセット 4 において主効果の有意傾向が見られた ($F(4, 35) = 2.186$, $p < .1$) が，データセット 1，データセット 2，データセット 3 では有意差が認められなかった ( $F(4, 26) = .395$, n.s.; $F(4, 25) = .543$, n.s.; $F(4, 40) = .808$, n.s.; ) ．

各データセットにおける correct 水準における，post 時期の正答率と成長度合いの関係をを図に示す（データセット 1: 図 5.7, データセット 2: 図 5.8, データセット 3: 図 5.9, データセット 4: 図 5.10 ）．これらの図の横軸は変更度合いの各段階を，縦軸は成長度合いを表している．

それぞれのデータセットでの結果において，回答変更率により成長度合いに差があるかを検証するために，独立変数を rate，従属変数を成長度合いとする一要因の分散分析を行った．その結果，全てのデータセットにおいて主効果の有意差が認められなかった ( $F(4, 26) = .024$, n.s.; $F(4, 25) = .378$, n.s.; $F(4, 40) = .215$, n.s.; $F(4, 35) = .559$, n.s. ) ．

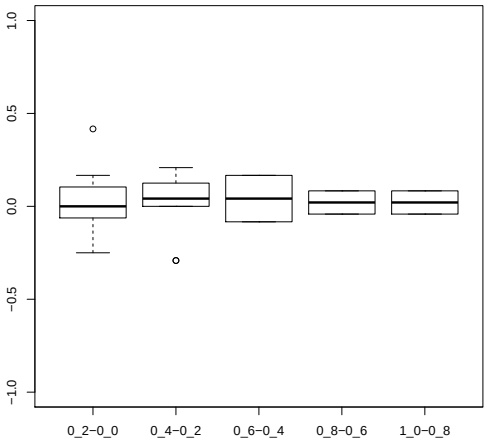


図 5.7　データセット 1 の correct 条件における回答変更率と成長度合い

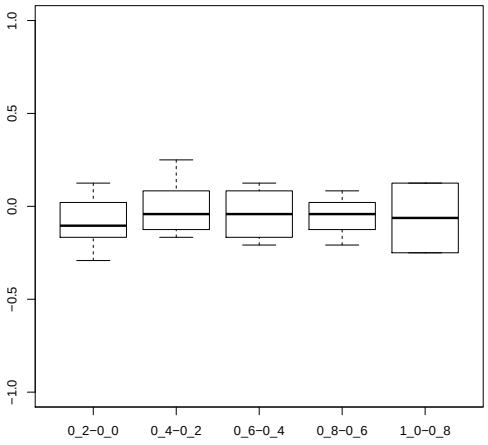


図 5.8　データセット 2 の correct 条件における回答変更率と成長度合い

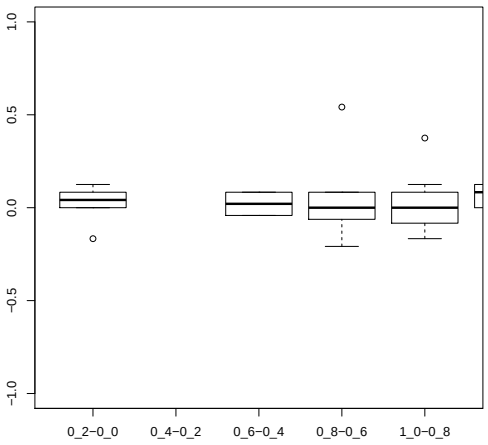


図 5.9　データセット 3 の correct 条件における回答変更率と成長度合い

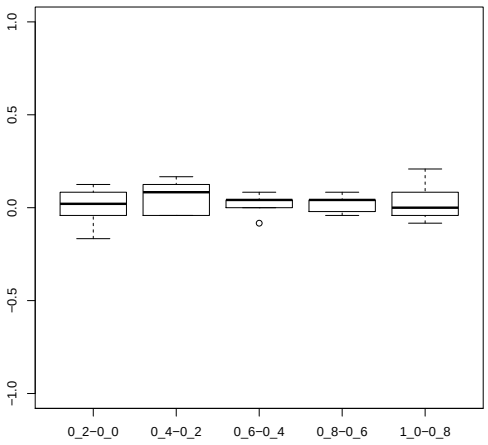


図 5.10　データセット 4 の correct 条件における回答変更率と成長度合い

### 5.1.4　考察

データセット 3 およびデータセット 4 を用いた実験では，テスト時期要因の主効果が認められた．一方で参考回答要因と交互作用については有意差が認められなかった．このことから，これらの実験においては自己補正による学習の転移は見られなかった．データセット 3 およびデータセット 4 でのテスト時期要因の主効果に関しては，タスクの繰り返しによる正答率の改善であると考えられる．テスト時期要因の主効果はデータセット 1 およびデータセット 2 では見られないことから，平均正答率が高くなるような課題の方が繰り返しによる正答率の改善が生じやすいと考えられる．

# 第 6 章

# 結論

## 6.1 総合考察

クラウドソーシングにおいて，ワーカから得られる成果物の品質を管理することは重要な研究課題の 1 つである．これまでに多くの研究がこの課題に取り組んでおり，本研究ではその 1 つである Shah らが提案した自己補正に着目した．自己補正はタスク結果の品質を改善することを目的とした手法であり，1 つのタスクに 2 度の回答の機会を与えることで，ワーカ自身が自分の誤答を補正できるのが特徴である．自己補正は，多数決をはじめとするタスク結果の集約手法や，優れたワーカの検出やタスク割り当て手法と組み合わせることが容易であることから，多くの場面で活用できる可能性がある．さらに，ワーカ自身の自らの誤答を気づかせる機会を与えることが，ワーカのその後の回答品質の改善にも効果があると期待できる．しかし自己補正が提案された論文では，シミュレーションによる評価のみが行われており，現実のクラウドソーシングでのワーカにもシミュレーションと同様の効果が得られるかは不明であった．

そこで本研究では，現実のクラウドソーシングワーカが自己補正を適用したタスクに取り組む実験により，自己補正の効果について検討した．実験 1A では自己補正の短期的な効果と長期的な効果を鳥の課題で検討した．実験 1B では，実験 1A で見られた効果が，別の難しいタスクにおいても同様の傾向であるかを，画家分類課題を用いて検討した．実験 2 では，特に実験 1A で見られた長期の効果が，類似した別の課題においても品質の改善をもたらすかを複数の画像分類課題を組み合わせて検討した．

## 6.2 自己補正の短期的効果

実験 1A と実験 1B での自己補正のステージ要因と正答率の分析から，次の結果が得られた．まず，自己補正をタスクに適用することで，現実のクラウドソーシングワーカから得られるタスク結果の品質が改善されることがわかった．実験 1A では With reference 条件にて有意な正答率の改善が見られたことから，自己補正における参考回答の提示は，タスク結果の品質改善のための重要な要素であることが示唆された．さらに，実験 1B では correct 条件にて有意な正答率の改善が見られたことから，参考回答の品質は重要であると考えられるが，参考回答の品質と自己補正によるタスク結果の改善の関係は，これらの実験からは不明である．

実験 1B の random 条件においても，ステージ 2 の成績がステージ 1 を下回る傾向は見られなかったため，何らかの手法に基いて参考回答を提示できる場合には，参考回答を提示することが有効であると考えられる．ただし，参考回答の内容や提示の方法は，ワーカがタスクに継続して取り組む際の動機づけを左右する要因になると考えられるため注意が必要である．

Shah らはワーカが自己補正により真面目に取り組むための報酬アルゴリズムが，今回は作

業を終えたワーカに対して定額の報酬を支払った．それにもかかわらず，タスク結果の品質改善が見られたことから，自己補正は独自の報酬アルゴリズムを導入することが難しい状況 (例えばワーカに対して一定の報酬を支払うことにのみ対応しているサービスを用いる場合など) においても有効な手法であると言える．

## 6.3 自己補正の長期的効果

実験 1A の結果から，ワーカが自己補正に連続で取り組むことで，ワーカ自身の回答品質の改善につながることが示唆された．また，回答品質の改善はテスト時期の pre-mid 間よりも mid-post 間で大きくなることから，改善にはある程度のタスク数が必要であることが分かる．ただし，今回の実験からはワーカの学習に必要なタクス数は自明でなく，これは各ワーカの状態や扱う課題などの要因に左右されると考えられる．

さらに実験 1B の結果から，自己補正に連続で取り組んだとしても，全体の傾向としてワーカ自身の回答品質の改善に繋がらない例があることが示された．実験 1B では絵画の画像を提示してその作者を推定する課題を扱ったが，全体を通して平均正答率が低く，学習効果も見られなかった．実験 1B では実験 1A よりも多くの学習タスクを割り当てたが，扱う課題によっては学習を促すことが難しいことが分かった．同様の課題についてより多くの学習タスクを割り当てることで，学習効果が見られる可能性は否定できない．ただし，ワーカが継続してタスクにより組みやすくするための支援が必要であると考えられ，例えば継続してタスクに取り組むことに対する報酬を与えるなどが挙げられる．

実験 1A, 実験 1B を通して，全体の傾向にかかわらず，一部のワーカは pre から post にかけて正答率が改善することを確認することが出来た．すべてのワーカが高い学習意欲を持つとは考えにくいため，学習効果が見られたワーカに注目して手法の評価をしたり，彼らを早期に発見する技術が重要である．

## 6.4 学習の転移について

実験 2 では，自己補正タスクによる学習の転移が生じるかを明らかにする実験を検討したが，自己補正の要因については有意差が認められなかった．自己補正タスクにおけるタスク設計やタスク割り当て，扱う課題の工夫などにより学習の転移を促すことができれば，クラウドソーシングプラットフォームにおけるワーカの育成に貢献できる可能性があることから検証する価値があると考えられる．

## 6.5 正答率が改善したワーカの分析

複数の実験結果の分析から，弱い傾向ではあるものの，自己補正における回答の変更頻度と post テストの正答率には関連があることが示唆された．ただし，回答の変更頻度と post テストの成績および成長度合いの分布は，実験ごとに大きく異なることから，より多くの実験的な検証が必要である．

## 6.6　今後の課題

### 6.6.1　自己補正タスクの繰り返しについて

本研究の実験では，自己補正タスクを数十回連続で割り当てることにより，ワーカの正答率が改善されることを示した．ただし，学習に必要な繰り返し数については明らかではない．十分な学習に必要な繰り返し数は，ワーカや取り扱う課題などの要因により変化することが予想される．

### 6.6.2　画像分類課題以外への応用

本研究では，すべての実験において画像分類課題による実験を行った．一方で，画像分類課題以外の課題に対して自己補正を適用した場合にも同様の傾向が見られるかは注目すべき課題である．

### 6.6.3　インセンティブ設計との組み合わせ

本研究では，すべての実験において定額の報酬をワーカに対して支払った．かかわらず自己補正の短期的・長期的効果が観察されたことから，動的な報酬の変更が困難であるようなクラウドソーシングプラットフォームに置いても自己補正が有効であると考えられる．一方で，Shah らは自己補正を提案した論文にて，自己補正のための報酬設計アルゴリズムを提案している [4]．この報酬アルゴリズムを始めとする，既存の報酬アルゴリズムを自己補正に適用した場合の，短期的および長期的効果を検証することは興味深い課題の 1 つである．

報酬アルゴリズムを導入することにより，ワーカに対してより多くの学習タスクを割り当てたり，単に他者回答へと変更することを抑制することができると考えられる．

### 6.6.4　参考回答の選び方

実験 1A では，自己補正のステージ 2 で提示する参考回答として，課題の正答率に基づいて信頼性の高いワーカを選択肢，彼らの回答を用いた．実験 1B では，参考回答として正答やランダムな回答を用いた．しかし，現実の正解が未知である課題を扱うクラウドソーシングにおいて，信頼性の高い回答を得ることは困難であることが多い．クラウドソーシングではゴールドスタンダードクエスチョンや多数決の結果などに基づいて信頼性の高い回答を得る方法が広く採用されていることから，自己補正における参考回答においてもこれらの方法を応用することが可能であると考えられる．

松原らは [27]，自己補正の参考回答として機械学習に基づく分類器の推論結果を提示することを試みた．参考回答として提示するための回答として推論結果を用いることは，別のワーカから回答を得るよりも低いコストで実現できるため現実的な手段の 1 つであると考えられる．

## 6.7　まとめ

本研究では，現実のクラウドワーカにおける自己補正の効果を明らかにするために，現実のクラウドワーカが自己補正タスクに取り組む実験を検討した．そして，自己補正がもたらす効果について，タスクの正答率や反応時間などを分析した．

実験 1A では，鳥の画像分類タスクに自己補正を適用し，自己補正で提示する参考回答の有

無を比較する実験を行った．実験 1B では，絵画の分類タスクに自己補正を適用し，自己補正で提示する参考回答の品質を比較する実験を行った．これらの実験結果から，自己補正によって現実のクラウドワーカがタスク結果を改善できることが示唆された．実験 1A では参考回答を提示する条件に，実験 1B では参考回答として正答を提示する条件において有意な改善が見られたことから，自己補正によるタスク結果の品質改善において参考回答は重要な要素であると考えられる．さらに，ワーカが自己補正を繰り返すことで，ワーカ自身の回答品質が改善される可能性が示唆された．この傾向は実験 1A では条件間に有意差が認められたのに対し，実験 1B では条件間の有意差が認められなかったことから，取り扱う課題などの要因に左右されやすいと考えられる．一方で，実験 1B のタスク結果について，自己補正タスクでの回答変更率を分析したところ，回答変更率が高すぎず低すぎないワーカの post テストの成績および成長度合いが高い傾向があることが示唆された．

実験 2 では，実験 1A および実験 1B で見られた自己補正によるワーカの品質改善が，学習課題と評価課題が異なる場合にも同様の傾向が見られるかを調べたところ，有意な傾向は認められなかった．

以上の実験結果から，自己補正がクラウドソーシングにおける品質改善に対して，タスク結果の改善とワーカの能力改善という 2 つの側面から貢献できる可能性を示した．

# 謝辞

本研究に取り組むにあたり，多くの方々からのご指導とご支援をいただきました．

森嶋厚行先生には，主指導教員として終始熱心なご指導を賜りました．ご多忙にもかかわらず日頃から相談にのっていただき，研究活動だけでなく学生生活や進路などに関しても多くのご助言をいただきました．深く感謝申し上げます．

森田ひろみ先生には，副研究指導教員を快く引き受けていただき，研究活動を通して実験計画や実験結果の解釈などについて多くのご指導とご助言をいただきました．心より感謝申し上げます．

ヤフー株式会社の清水伸幸様には，森嶋厚行先生との共同研究者として，本研究に関しても多くのご指導およびご助言をいただきました．心より感謝申し上げます．

融合知能デザイン研究室の松原正樹先生，渡辺知恵美先生には，研究室生活やゼミを通して丁寧なご指導とご助言をいただきました．そして，同研究室の学生の皆様，そしてスタッフの皆様にも日頃からお世話になりました．

杉本永森研究室および阪口研究室の先生方と学生の皆様には，合同ゼミを通して研究に関する多くのご助言をいただきました．

本研究の実験では多くのクラウドワーカの皆様にご協力をいただきました．彼らの協力により，この研究活動はより充実したものとなりました．ここに感謝いたします．

最後に，応援していただいた家族と友人の皆様に感謝申し上げます．

# 参考文献


[1] Eleanor Jack Gibson. Principles of perceptual learning and development. 1969.

[2] James J Gibson and Eleanor J Gibson. Perceptual learning: Differentiation or enrichment? *Psychological review*, Vol. 62, No. 1, p. 32, 1955.

[3] Everett Mettler and Philip J Kellman. Adaptive response-time-based category sequencing in perceptual learning. *Vision research*, Vol. 99, pp. 111–123, 2014.

[4] Nihar Shah and Dengyong Zhou. No oops, you won't do it again: Mechanisms for self-correction in crowdsourcing. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, Vol. 48 of *Proceedings of Machine Learning Research*, pp. 1–10, New York, New York, USA, 20–22 Jun 2016. PMLR.

[5] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Comput. Surv.*, Vol. 51, No. 1, pp. 7:1–7:40, January 2018.

[6] Nguyen Quoc Viet Hung, Nguyen Thanh Tam, Ngoc Tran Lam, and Karl Aberer. An evaluation of aggregation techniques in crowdsourcing. In *Web Information Systems Engineering - WISE 2013 - 14th International Conference, Nanjing, China, October 13-15, 2013, Proceedings, Part II*, pp. 1–15, 2013.

[7] Srikanth Jagabathula, Lakshminarayanan Subramanian, and Ashwin Venkataraman. Reputation-based worker filtering in crowdsourcing. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pp. 2492–2500. Curran Associates, Inc., 2014.

[8] Lora Aroyo and Chris Welty. Measuring crowd truth for medical relation extraction. In *AAAI 2013 Fall Symposium on Semantics for Big Data*, 2013.

[9] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. Toward a learning science for complex crowdsourcing tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 2623–2634. ACM, 2016.

[10] Peter Kinnaird, Laura Dabbish, Sara Kiesler, and Haakon Faste. Co-worker transparency in a microtask marketplace. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pp. 1285–1290. ACM, 2013.

[11] Gary Hsieh and Rafał Kocielnik. You get who you pay for: The impact of incentives on participation bias. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pp. 823–835. ACM, 2016.

[12] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI 2017)*. ACM âĂŞ

Association for Computing Machinery, May 2017.

[13] Ryan Drapeau, Lydia B Chilton, Jonathan Bragg, and Daniel S Weld. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*, 2016.

[14] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pp. 1013–1022. ACM, 2012.

[15] Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, and Stefan Dietze. Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks. *ACM Transactions on Computer-Human Interaction (TOCHI)*, Vol. 24, No. 4, p. 30, 2017.

[16] Masayuki Ashikawa, Takahiro Kawamura, and Akihiko Ohsuga. Proposal of grade training method in private crowdsourcing system. In *Third AAAI Conference on Human Computation and Crowdsourcing*, 2015.

[17] Ryo Suzuki, Niloufar Salehi, Michelle S. Lam, Juan C. Marroquin, and Michael S. Bernstein. Atelier: Repurposing expert crowdsourcing tasks as micro-internships. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pp. 2645–2656, New York, NY, USA, 2016. ACM.

[18] Azad Abad, Moin Nabi, and Alessandro Moschitti. Autonomous crowdsourcing through human-machine collaborative learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pp. 873–876, New York, NY, USA, 2017. ACM.

[19] Edith Law, Ming Yin, Joslin Goh, Kevin Chen, Michael A. Terry, and Krzysztof Z. Gajos. Curiosity killed the cat, but makes crowdwork better. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pp. 4098–4110, New York, NY, USA, 2016. ACM.

[20] Nguyen Quoc Viet Hung, Duong Chi Thang, Matthias Weidlich, and Karl Aberer. Minimizing efforts in validating crowd answers. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pp. 999–1014. ACM, 2015.

[21] Daniel Haas, Jason Ansel, Lydia Gu, and Adam Marcus. Argonaut: macrotask crowdsourcing for complex data processing. *Proceedings of the VLDB Endowment*, Vol. 8, No. 12, pp. 1642–1653, 2015.

[22] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1631–1640. ACM, 2015.

[23] Hyun Joon Jung and Matthew Lease. Modeling temporal crowd work quality with limited supervision. In *Third AAAI Conference on Human Computation and Crowdsourcing*, 2015.

[24] Manas Joglekar, Hector Garcia-Molina, and Aditya Parameswaran. Evaluating the crowd with confidence. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pp. 686–694, New York, NY, USA, 2013. ACM.

[25] Akash Das Sarma, Aditya Parameswaran, and Jennifer Widom. Towards globally

optimal crowdsourcing quality management: The uniform worker setting. In *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD '16, pp. 47–62, New York, NY, USA, 2016. ACM.

[26] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[27] Masaki Matsubara, Masaki Kobayashi, and Atsuyuki Morishima. A learning effect by presenting machine prediction as a reference answer in self-correction. *In Proceedings of The Second IEEE Workshop on Human-in-the-loop Methods and Human Machine Collaboration in BigData (IEEE HMData2018)*, pp. 3521–3527, 2018.