DEIM Forum 2018 G6-3

クラウドワーカの品質改善における他者回答提示の短期的・長期的効果

小林　正樹† 　松原　正樹†† 　森田ひろみ†† 　清水　伸幸††† 　森嶋　厚行††

† 筑波大学 図書館情報メディア研究科　〒 305-8550 茨城県つくば市春日 1-2
†† 筑波大学 図書館情報メディア系　〒 305-8550 茨城県つくば市春日 1-2
††† ヤフー株式会社　〒 102-8282 東京都千代田区紀尾井町 1-3 東京ガーデンテラス紀尾井町 紀尾井タワー
E-mail: †s1311495@klis.tsukuba.ac.jp, ††{masaki,morita,mori}@slis.tsukuba.ac.jp, †††nobushim@yahoo-corp.jp

あらまし　クラウドソーシングタスクにおける自己補正は, ワーカが他のワーカの回答を参照することで, 自身の回答を再評価するための仕組みである. 自己補正は同じ質問に対して複数のワーカが取り組む状況において, 低コストで高品質なタスク結果を得るのに有効であることが，過去の研究においてシミュレーションで示されている．しかし, 自己補正が現実のワーカに対して有効であるかは明らかでない. 本論文では, 自己補正の効果を評価するために, 現実のクラウドワーカによる実験を実施した. 実験の結果から次の 2 点が明らかとなった. (1) クラウドソーシングにおける自己補正が，現実のクラウドワーカに対しても有効であること (2) 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にも改善が見られることがあること

キーワード　クラウドソーシング

1. はじめに

クラウドソーシングにおいて，群衆から得られた成果物の品質を保証することは重要な課題の 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きた．

成果物の品質を高めるための基本的な戦略は，信頼できるワーカに対してタスクを割り当てることである．具体例として，Amazon Mechanical Turk における MTurk Master Worker(注1) という仕組みが挙げられる．リクエスタは，タスクをワーカに依頼する際に追加の料金を支払うことで，Master の資格を持つワーカに優先してタスクを割り当てることが出来る．ワーカはプラットフォームが定めた基準を満たすことで，マスターの資格が与えられ，マスターに対して割り当てられたタスクに取り組むことが出来るようになる．この仕組みにより，リクエスタは信頼できるワーカから高品質な回答を得ることが出来るのである．ただし，マスターの資格を持つワーカの数は限られているため，大量の作業を目的の期間内に終えるといった要求に答えられないことがある．

このような場合，ワーカはそれぞれが異なる品質の回答をもたらすことが想定される．このような状況で成果物の品質を保証するために用いられる一般的な手法として，複数のワーカから得られた回答の多数決が挙げられる．タスクに対する回答を複数のワーカから集め，それらを集約することで回答の品質を高めるのがねらいである．

ただし，これらの手法はその時点での回答の品質を改善するものであり，それ以降のタスク結果の品質改善をもたらすことはない．ワーカがもたらす回答は，クラウドソーシングにおける成果物の品質を左右する重要な要因であるため，ワーカの回答能力を向上させることが重要である．高品質な回答をもたらすワーカは常に求められている．未熟なワーカが通常の作業に取り組むことによって，回答品質の改善を促すことができれば，ワーカとリクエスタそれぞれにとって好ましい状況となる．

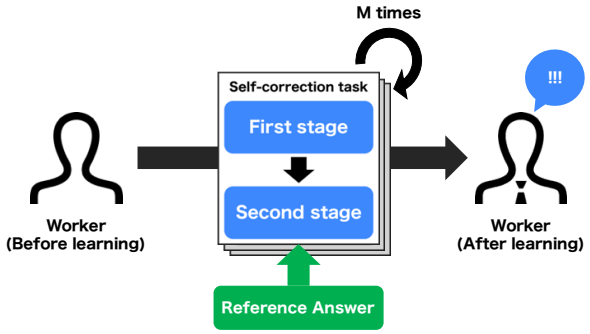

図 1　本研究の概要図：自己補正は，ワーカが質問に回答 (First stage) した後に，自身の回答と参考回答を見て修正する作業 (Second stage) を行うことで，タスク結果の品質改善を図る手法である．実験では，自己補正タスクを現実のワーカに割り当てることで，自己補正の有効性を明らかにするとともに，学習においても効果があるかを明らかにする．

タスク結果の品質を改善する手法の 1 つに，Nihar [1] らが提案したクラウドソーシングタスクにおける自己補正がある．自己補正を適用したタスクでは，ワーカは 1 つのタスクに対して 2 回の回答する機会を与えられる．自己補正の重要な要素は，2 回目の回答の際に，ワーカ自身の 1 回目の回答と，既に同じ質問に回答した別のワーカの回答を与えられた上で，最終的な回答を判断することである．ワーカが自身の誤りを訂正する機会を提供することで，タスクの金銭的なコストとタスクの完了までに必要なワーカの数の削減が期待できる．

さらに，自己補正では事前にラベル付されたデータセットを用いてワーカを訓練する過程を必要としない．そのため，商用のクラウドソーシングプラットフォームに掲載するタスクに

(注1)：https://www.mturk.com/worker/help#what_is_master_worker

対して容易に適用することが出来る．多数決を始めとするワーカの回答を集約する手法などと組み合わせることにより，タスク結果の品質改善についてより大きな効果も期待できる．一方で，Nihar らの論文ではその有効性がシミュレーションのみで示されており，現実のクラウドワーカにおいても自己補正が有効であるかは明らかでない．自己補正については，本稿の 3. 節でその詳細を述べる．

本稿では，自己補正について現実のクラウドワーカを用いた実験を行うとともに，自己補正の長期的な効果についても評価する．長期的な効果とは，ワーカが自己補正を繰り返した際の学習効果のことである．知覚学習においては，ある事象に関する知覚を繰り返して経験することで，知覚に関する成績や反応時間が向上することが知られている [2] [3]．これまでに多くの研究によって，被験者に対するフィードバックの与え方や頻度などを工夫することで，学習後の成績の向上や学習効率の改善に繋がることが示されている [4] [5]．自己補正で他者回答の提示することは，知覚学習の分野におけるフィードバックの一種に相当すると考え，自己補正を繰り返すことで繰り返し学習の効果が見られると考えた．

実験では，現実のクラウドワーカにおける自己補正の有効性の検証に加えて，自己補正の繰り返しによるワーカの回答品質の改善にも注目する (図 1)．実験 1 では参考回答の有無を，実験 2 では参考回答の品質が自己補正の効果に与える影響を比較する．本論文の貢献は次のとおりである．

（1） クラウドソーシングにおける自己補正が現実のクラウドワーカに対しても有効であることを示す

（2） ワーカに自己補正を連続で与えることで，ワーカ自身の回答品質にも改善が見られるかを明らかにする

2. 関 連 研 究

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課題であり，これまでに多くの研究がこの問題に取り組んできた．

回答の品質を改善するために，ワーカの能力を向上する場合，広く検討されているのはワーカが本番のタスクに取り組む前に，訓練のための作業に取り組んでもらう方法である．訓練タスクを終えた後に本番のタスクに取り組むことで，ワーカから得られる回答の品質が改善されることが知られており，効率的な学習を促すためのタスク割り当て手法 [6] などが提案されている．このようなアプローチを用いる場合，回答が既知のタスクを十分に用意する必要がある．

別のアプローチとして，1 つのタスクを複数のワーカに割り当て，複数の結果を集約する方法がある．複数の結果を多数決などの方法により集約することで，一部のワーカが誤った回答をした場合でも，全体としては品質の高い回答を得ることが出来るのである．多数決は様々な文脈で用いられる手法であるが，クラウドソーシングの文脈においてはワーカごとの性質や，回答の傾向などの特徴を活用した応用例が提案されている [7] [8] [9] [10].

この 2 つのアプローチとは対象的に，本研究では訓練のためのデータセットを用意することが難しい状況において，自己補正によってワーカの回答品質を改善することで，成果物の品質を改善しようとする点に独自性がある．自己補正では，他者の回答を提示することでタスク結果の品質改善を試みる手法であるが，同様にワーカに対して別のワーカによる評価結果や回答の理由，回答の傾向など与えることでタスク結果の品質改善をする手法が提案されている [11] [12] [13] [14]．一方で，作業に取り組むワーカに対して，同様の作業に多くのワーカが関わっていることを知らせることが，ワーカの作業に対する動機づけを低下させることが報告されており [15]，このような情報の提示方法はワーカの動機づけを左右する要因であるといえる．ワーカの作業に対する動機づけについては，作業に対する対価が成果物の品質を左右することが知られている [16].

このような参考回答をもたらすワーカを選択したり，ワーカに対する評価を決定する上で重要となるのがワーカの能力を測定する手法である [17] [18]．ワーカの品質を評価するための基本的な手法は，ワーカが取り組むタスクの一部に正答が既知のタスクを含めておき，それらの正答率を算出する方法である．評価の正確性を高めるためには，ワーカがタスクに取り組み始めた直後に評価するのではなく，継続的に評価を行うことが重要であることが報告されている [19]．自己補正の第 2 段階で提示する回答には，既に同じタスクに回答したワーカの回答を用いることが考えられるが，ワーカの選び方についてこれらの手法を用いることが出来る．

3. 自 己 補 正

この節では，Nihar 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 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．

3.2 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が存在することが想定される．そこで，Nihar 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 タスクで正答することが最も価値が高く，第 2 段階で他者の回答を支持することは価値が低いような設定となっている．

3.3 シミュレーション

Nihar らは，自己補正の有効性を明らかにするために，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

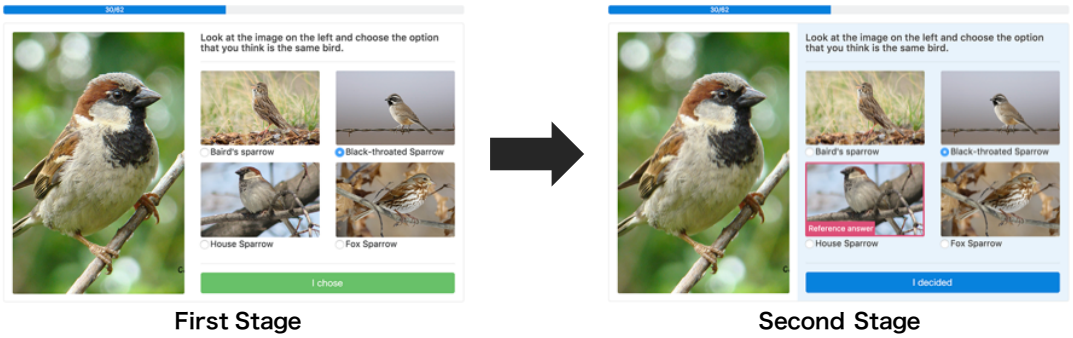

図 2　自己補正タスクの例

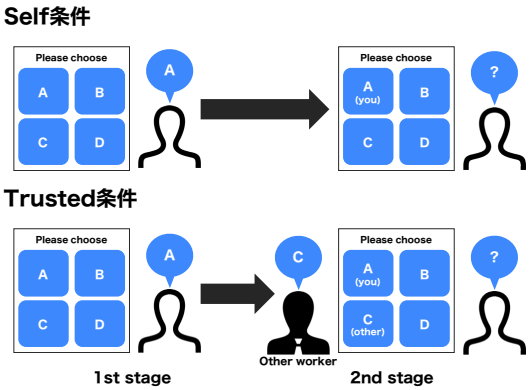

図 3　実験 1 で比較する参考回答の条件

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．

4. 実　　験 1

実験 1 では，前節で述べたクラウドソーシングにおける自己補正について，

（1） 現実のクラウドワーカにおいても短期的なタスク結果の品質改善が見られるか

（2） 自己補正を繰り返すことが，ワーカ自身の長期的な回答品質の改善に繋がるか

を明らかにすることを目的とした実験を行う．本実験は筑波大学図書館情報メディア系研究倫理審査委員会の承認を得ている．実験の概要を図 3 に示す．

4.1 実験参加者

Yahoo!クラウドソーシング (注2) 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，クラウドワーカ 200 名が参加した．参考回答の有効性を調べるために，実験参加者のうち 100 名を参考回答ありのグループ，別の 100 名を参考回答なしのグループとした．実験に最後まで参加した被験者には，回答の品質を問わず 100 円相当の報酬を支払った．

(注2)：https://crowdsourcing.yahoo.co.jp/

4.2 タ　ス　ク

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 [20] からを用いた．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．

4.3 実験の流れ

実験参加者は与えられた Web ページに提示される 92 個のタスクを順に回答する (表 1)．タスクは 3 回のテストフェーズ (pre, mid, post-test) と 2 回の学習フェーズ (learn1, learn2) で構成されている．テストフェーズではワーカの能力を測定するためのタスクが 12 個提示される．テストフェーズは学習フェーズの前後で割り当てられ，テストフェーズの成績の変化を学習の効果として扱う．

学習フェーズでは，自己補正を適用したタスクが 28 種類ほど表示される．自己補正を適用したタスクにて参考回答が与えられるか否かは，実験参加者が割り当てられたグループにより決定する．自己補正の第 1 段階の回答と第 2 段階の回答を比較することで，自己補正の効果を明らかにする．

参考回答ありのグループで提示される参考回答には，参考回答なしのグループの回答を用いた．参考回答なしのグループの

表 1　実験 1 の流れとフェーズ設定

Phase	Name	Task type	Task number
1	Pre-test	Test	12
2	Learn1	Self-correction	28
3	Mid-test	Test	12
4	Learn2	Self-correction	28
5	Post-test	Test	12

表 2　実験 1 の結果の概要 (pre-test の成績)

条件	フィルタ	N	平均	標準偏差	min	max
Trusted	None	98	0.816	0.132	0.250	1.0
Trusted	Under 25%	86	0.824	0.134	0.417	1.0
Self	None	98	0.825	0.147	0.250	1.0
Self	Under 25%	84	0.831	0.136	0.333	1.0

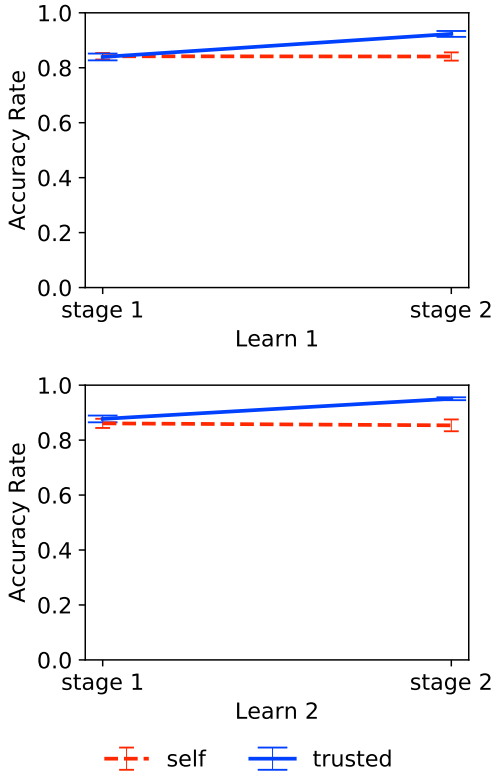

図 4 （実験 1） ステージ要因と参考回答の関係

うち，成績の優れていた参加者を 20 名を選び，彼らの回答を参考回答とした．参考回答ありの条件では，自己補正の第 2 ステージにおいて，他者の回答が赤枠で示される．

4.4 結　　果

実験参加者の人数と pre-test の成績を表 2 に示す．今回の実験ではワーカはタスクに連続で取り組む必要があるため，タスクの途中からランダムな回答をするようなワーカが見られた．そこで，mid-test, post-test の成績が 25%を下回るようなワーカについては以降の集計から除外した．

4.4.1 短期的効果

自己補正についてワーカに対する短期的な効果を評価する．learn1 および learn2 における，自己補正の第 1 段階と第 2 段階での正答率の変化を図 4 に示す．参考回答および自己補正のステージによってタスクの正答率の差があるかを検証するために，独立変数を参考回答と自己補正のステージ，従属変数をタスクの正答率とする 2 要因の分散分析を行った．ステージ要因については learn1 と learn2 の正答率の平均値を用いた．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) = 10.454, p <.001; F(1, 168) = 39.321, p <.001; F(1, 168) = 48.290, p <.001) .

交互作用が見られため，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を行ったところ，参考回答が trusted の条件では有意な単純主効果が認められた (F(1, 168) = 88.42, p <.001) が，self の条件では有意でなかった (F(1, 168) = .23, ns) ．さらに，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を行ったところ，ステージが 1st の条件では有意な単純主効果が認められなかった (F(1, 168) = .18, ns) が 2nd の条件では有意であった (F(1, 168) = 31.82, p <0.001).

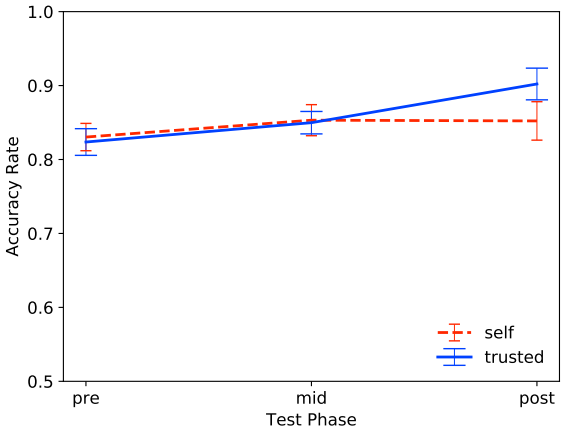

図 5 （実験 1） テスト時期と参考回答の関係

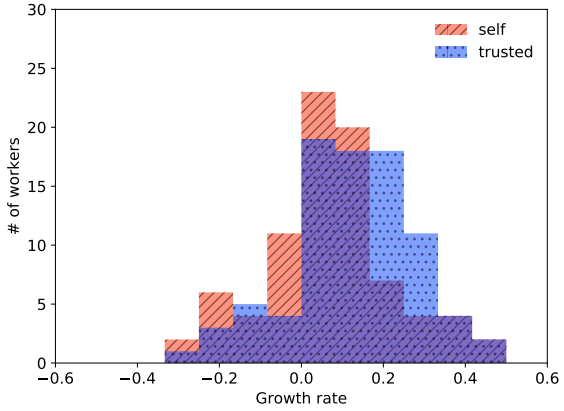

図 6 （実験 1） 成長度合いの分布

4.4.2 長期的効果

自己補正についてワーカに対する長期的な効果を評価する．テスト時期と参考回答の関係を図 5 に示す．参考回答およびテストの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の時期，従属変数をタスクの正答率とする 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) = 8.731, p <.001; F(2, 336) = 3.5, p <.05;) が，参考回答要因の主効果は有意でなかった (F(1, 168) = 0.635, ns).

交互作用が有意のため，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，参考回答が trusted の条件では有意な単純主効果が認められた (F(2, 336) = 11.19, p <.001) が，self の条件では有意でなかった (F(2, 336) = 1.14, ns) ．さらに，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，テスト時期が pre と mid の条件では有意な単純主効果が認められなかった (F(1, 168) = .1, ns; F(1, 168) = .03, ns) が，post の条件では有意であった (F(1, 168) = 4.48, p <.05) .

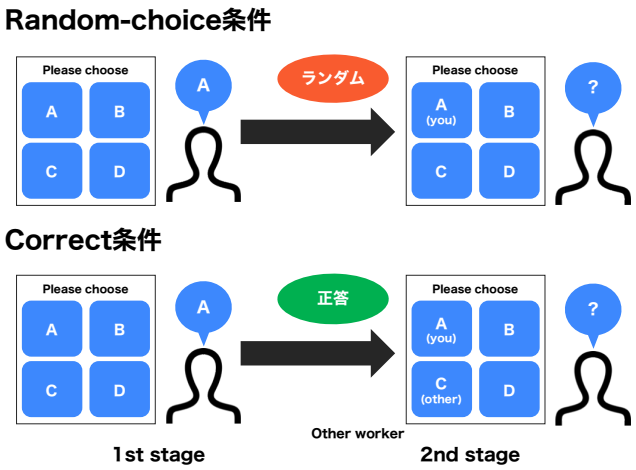

図 7　実験 2 で比較する参考回答の条件

4.5　考　　察

4.5.1　短期的効果

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は，参考回答が trusted の条件でのみ認められた．そして，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は，ステージが 2nd の条件でのみ認められた．このことから，高品質な回答をもたらすワーカの回答を参考回答として与えることで，自己補正によるタスク結果の品質改善が生じることが分かった．

4.5.2　長期的効果

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は，参考回答が trusted の条件でのみ認められた．そして，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は，テスト時期が post の条件でのみ認められた．このことから，trusted 条件の参考回答を提示する自己補正を繰り返すことで，ワーカ自身の回答品質が向上することが分かった．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．

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 6 に示す．参考回答が trusted 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，self の条件よりも多いことが分かる．このことから，trusted 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．

表 3　実験 2 でワーカが取り組む作業の流れ

Phase	Name	Task type	Task number
1	Pre-test	Test	12
2	Learn1	Self-correction	52
3	Mid-test	Test	12
4	Learn2	Self-correction	52
5	Post-test	Test	12

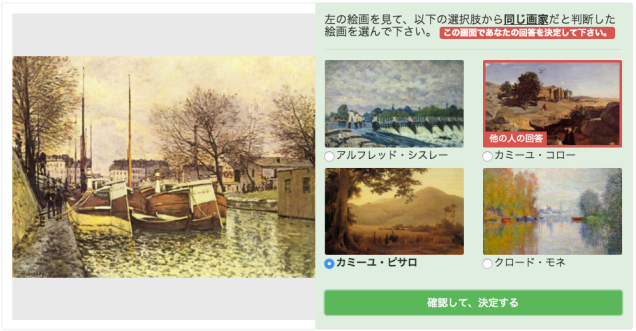

図 8　（実験 2）自己補正の第 2 段階の例

5.　実　　験 2

実験 2 では，課題の難易度を複雑にした場合の自己補正の短期的・長期的効果を明らかにする．実験 1 よりも平均正答率が低くなるようなデータセットを用いてタスクを作成する．加えて，学習フェーズで割り当てるタスクの数を実験 1 よりも多い設定とする．更に，参考回答の条件は，(1) 常に正解，(2) 常にランダムの 2 種類とする．これは参考回答の品質が，ワーカに与える影響を確認するためである．実験の説明について，実験 1 と同様の項目については説明を省略する．実験の概要を図 7 に示す．

5.1　タ　ス　ク

実験 2 では，絵画の画像を提示してその作者を選択する課題を扱う．タスクの形式は実験 1 と同様に 4 択の選択式とする (図 8)．絵画の画像は wikiart.org(注3) から収集した 4 名の作家の画像を用いる．

5.2　参 考 回 答

全て正解の場合と全てランダムの場合のグループに分け，それぞれ 100 名の被験者が参加する実験を行う．ランダムな回答は擬似乱数を用いて決定した．

5.3　実験の流れ

ワーカが実験で取り組む作業の流れを表 3 に示す．ワーカが 5 つのフェーズで構成されたタスクに取り組む点は実験 1 と同様であるが，Learn フェーズでのタスク数が異なる．

5.4　ワーカのフィルタ

実験の途中から意図の無い回答をするようなワーカを分析から除外するために，Learn フェーズに選択肢に表示されている画像を質問とするタスクを 4 つ含めた．これらのタスクに正答できなかったワーカについては実験結果の分析から除外する．

表 4　実験 2 の結果の概要 (pre テストの成績)

条件	フィルタ	N	平均	標準偏差	min	max
Correct	None	115	0.356	0.145	0.083	0.75
Correct	Gold	100	0.363	0.145	0.083	0.75
Random-choice	None	76	0.352	0.16	0.083	1.0
Random-choice	Gold	61	0.361	0.165	0.083	1.0

(注3)：https://www.wikiart.org/

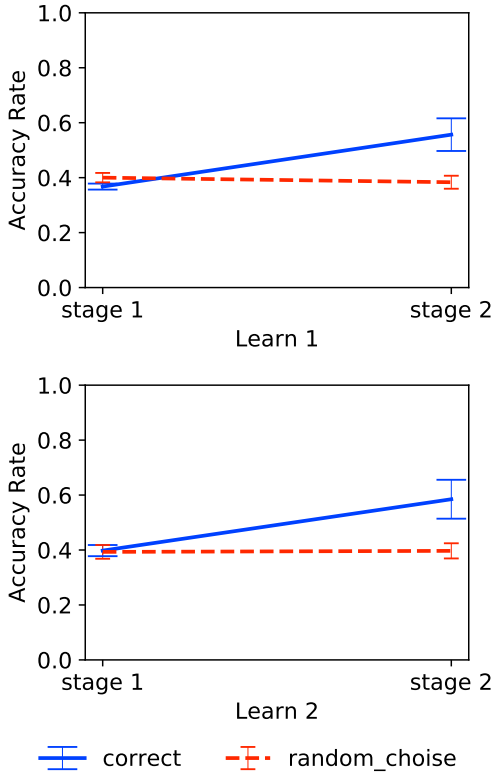

図 9 （実験 2） ステージ要因と参考回答の関係

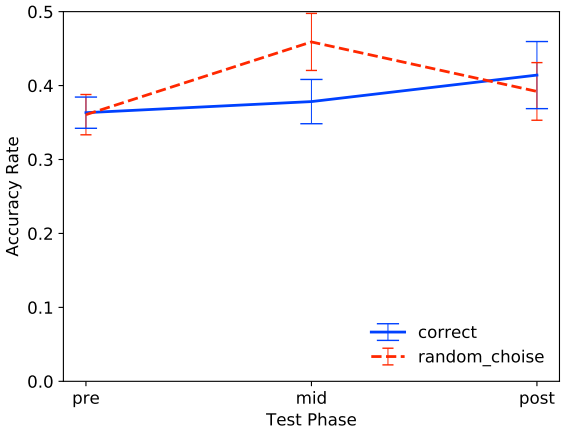

図 10 （実験 2） テスト時期と参考回答の関係

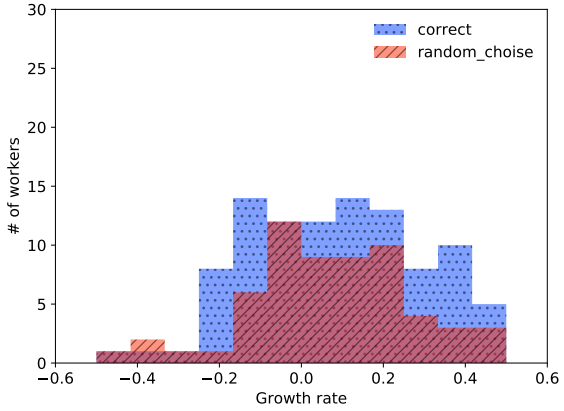

図 11 （実験 2） 成長度合いの分布

5.5 結　　果

実験 2 に参加した被験者の人数と pre テストの成績を表 4 に示す．

5.5.1 短期の効果

自己補正についてワーカに対する短期的な効果を評価する．learn1 および learn2 における，自己補正の第 1 段階と第 2 段階での正答率の変化を図 9 に示す．

参考回答および自己補正のステージによってタスクの正答率の差があるかを検証するために，独立変数を参考回答と自己補正のステージ，従属変数をタスクの正答率とする 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 ($F(1, 159) = 12.153$, $p < .01$; $F(1, 159) = 36.475$, $p < .001$; $F(1, 159) = 41.855$, $p < .001$) .

交互作用が有意のため，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を行ったところ，参考回答が trusted の条件では有意な単純主効果が認められた ($F(1, 159) = 103.25$, $p < .001$) が，self の条件では有意でなかった ($F(1, 159) = .07$, ns) ．さらに，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を行ったところ，ステージが 1st の条件では有意な単純主効果が認められなかった ($F(1, 159) = .52$, ns) が 2nd の条件では有意であった ($F(1, 159) = 26.63$, $p < 0.001$).

5.5.2 長期の効果

自己補正についてワーカに対する長期的な効果を評価する．テスト時期と参考回答の関係を図 10 に示す．参考回答およびテストの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の時期，従属変数をタスクの正答率とする 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 ($F(2, 318) = 6.213$, $p < .005$; $F(2, 318) = 5.399$, $p < .01$;) が，参考回答要因の主効果は有意でなかった ($F(1, 159) = 0.684$, ns).

交互作用が有意のため，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，参考回答が correct 条件と random の条件でそれぞれ有意な単純主効果が認められた ($F(2, 318) = 3.26$, $p < .05$ $F(2, 318) = 7.36$, $p < .005$) .

さらに，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，テスト時期が pre と post の条件では有意な単純主効果が認められなかった ($F(1, 159) = .1$, ns; $F(1, 159) = .43$, ns) が，mid の条件では有意であった ($F(1, 159) = 7.43$, $p < .05$) .

5.6 実験 2 の考察

5.6.1 短期の効果

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は，参考回答が correct の条件でのみ認められた．そして，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は，ステージが 2nd の条件でのみ認められた．この

ことから，参考回答として正答を与えることで，自己補正によるタスク結果の品質改善が生じることが分かった．実験 1 での短期の効果と同様の効果を，別の課題を用いた実験 2 でも確認することが出来たといえる．ただし，自己補正タスクの繰り返しの回数などの設定が異なることに注意しなければならない．

5.6.2 長期の効果

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は，参考回答が correct の条件と random の条件のそれぞれで認められた．そして，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は，テスト時期が mid の条件でのみ認められた．この結果は，参考回答として正答を提示することだけが，ワーカの回答品質の改善に必要な要素ではないことを示唆するものである．実験 2 のような平均正答率が低い課題においては，他者の回答として提示された内容に疑いを持ち，より注意深く回答するといった行動が想定される．また，random な参考回答を与えた場合の mid 時期の成績は，同条件の pre 時期や正答を提示する場合の同時期を上回る一方で，post 時期の正答率は減少していることから，タスクに連続で取り組むことで集中力が途切れたり，他者回答をそのまま採用するようなワーカが増えていることが想定される．

各ワーカの成長度合いについてのヒストグラムを図 11 に示す．僅かではあるが，参考回答が correct の場合に成長度合いが 0.2 から 0.4 に相当するワーカが存在することが分かる．このことから，一部のワーカについては自己補正によるワーカ自身の回答品質の改善が確認できた．

6. 考　　察

6.1 自己補正の短期的効果

実験 1 の結果から，Nihar らが提案したクラウドソーシングタスクにおける自己補正が，現実のクラウドワーカの回答品質の改善に対して有効であることが示された．Nihar 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．

実験 2 では，実験 1 よりも平均正答率が低くなるような課題を与える場合において，参考回答として正答を与える場合とランダムな回答を与える場合を比較した．その結果，正解を与える場合においては実験 1 と同様にタスク結果の品質改善の効果が見られた．ランダムな回答を提示した場合でもステージ 2 の成績がステージ 1 の成績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．

6.2 自己補正の長期的効果

実験 1 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 2 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 2 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 2 では実験 1 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 1, 実験 2 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．

7. まとめと今度の課題

本研究では，クラウドソーシングタスクにおける自己補正が，現実のクラウドワーカに対しても有効であるか，加えて自己補正を繰り返すことがワーカ自身の能力改善に繋がるかを明らかにした．その結果，(1) 自己補正が現実のクラウドワーカに対しても有効な手法であること，(2) 自己補正の繰り返しによるワーカ自身の能力改善は確認されたが，タスクの難易度などの要因に左右されることが示唆された．

今後の課題としては，難易度や性質の異なるデータセットを用いた実験や能力の改善が見込まれるようなワーカを早期発見する手法の検討が挙げられる．

謝　　辞

本研究の一部は JST CREST (#JPMJCR16E3) の支援による．

文　　献

[1] 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.

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

[3] Richard A Schmidt and Robert A Bjork. New conceptualizations of

practice: Common principles in three paradigms suggest new concepts for training. *Psychological science*, Vol. 3, No. 4, pp. 207–218, 1992.

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

[5] Nate Kornell and Robert A Bjork. Learning concepts and categories.

[6] 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.

[7] Nguyen Quoc Viet Hung, Nguyen Thanh Tam, Lam Ngoc Tran, and Karl Aberer. An evaluation of aggregation techniques in crowdsourcing. In *International Conference on Web Information Systems Engineering*, pp. 1–15. Springer, 2013.

[8] 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.

[9] Shunsuke Kajimura, Yukino Baba, Hiroshi Kajino, and Hisashi Kashima. Quality control for crowdsourced poi collection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 255–267. Springer, 2015.

[10] 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.

[11] 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.

[12] 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.

[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] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. ACM Association for Computing Machinery, May 2017.

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

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.