

國立臺灣大學電機資訊學院資訊網路與多媒體研究所
博士論文

Graduate Institute of Networking and Multimedia
College of Electrical Engineering and Computer Science
National Taiwan University
Doctoral Dissertation



設計解決複雜的創造性任務
Designing for Complex Creative Task Solving

黃怡靜

Yi-Ching Huang

指導教授：許永真博士
Advisor: Jane Yung-jen Hsu, Ph.D.

中華民國 107 年 6 月

June, 2018



國立臺灣大學博士學位論文

口試委員會審定書

設計解決複雜的創造性任務

Designing for Complex Creative Task Solving

本論文係黃怡靜君（學號 D00944010）在國立臺灣大學資訊網路與多媒體研究所完成之博士學位論文，於民國一百零七年六月四日承下列考試委員審查通過及口試及格，特此證明

口試委員：

許永真

（簽名）

（指導教授）

王浩全

階段盛

陳國寧

余朝豪

何建儒

梁容輝

古倫經

楊佳玲

所長：





誌謝

在博士研究的這段歲月，絕對是我人生中最難忘的經歷。我學會如何在一次又一次的失敗中重新站起，拍拍身上的灰塵，繼續堅定地向前。感謝每一次的失敗與歷練，讓我更加堅強。雖然過程中曾經迷失，也曾經無數次地懷疑自己，但幸運的是，在孤獨的奮鬥背後總有一群老師、朋友，以及我親愛的家人在身後支持著我，不斷的給我力量，讓我能夠克服任何困難。這段過雖然艱辛但也喜悅，感謝我所走過的每一條路，以及所遇到的寶貴的人、事、物，因為有你們，才有現在的我。

在學術與人生的路上，感謝影響我最深的指導教授許永真老師。我永遠都記得十年前大四下第一次跟你討論研究，天花亂墜的聊天，心中馬上就確定你是我要跟隨的指導老師，於是進入了人工智慧代理人實驗室 (iAgents Lab)，開啟我的研究生涯。碩士畢業後，工作了三年又回到實驗室繼續完成博士研究。從你身上，我不只學到了專業，還有謙卑又自信的態度，以及更多的人生道理。很高興在研究的路上，有你的指導與相伴。另外，我想感謝的是現在在美國 UC Davis 的王浩全老師，因為你多年的指導與提拔，讓我精進研究的思辨能力，更開啟了我在 CSCW 領域的研究，一起探索了許多有趣的想法與問題。

感謝陳玲鈴老師、梁容輝老師、徐宏民老師、余能豪老師、張俊盛老師、古倫維老師、何建儒老師、詹力韋老師以及張永儒老師，在我的博士論文提案以及學位考試給我許多寶貴的建議與指導。感謝泓其、啟嘉、Joey、以及過去與現在實驗室的學長姐與學弟妹們，也很感謝清大 CSC Lab 的學弟妹們，很開心能夠跟一起學習，並分享研究的喜

悅與辛苦。

感謝 OT 的好友們，讓我在追求研究的同時，也能繼續在科技藝術上有小小的貢獻，與你們聊天真的是一種享受，很感謝讓我遇見你們這群瘋狂的朋友們，你們的熱情總是給我滿滿的正向力量。另外，很感謝我的冰研死黨們：芭樂、驢子、劉嚕、頭頭，等高中死黨們，總是讓我想起小時候那股什麼都不怕的衝勁，即使遇到任何困難，總有你們在身旁邊吵吵鬧鬧，莫名什麼都克服了。

最後，感謝我最親愛的家人：媽咪、爸比、阿弟、哥哥，以及可愛的小咪醬與小花花。因為你們，我才能勇敢地追尋我的夢想。當然，還有小河童，因為有你的陪伴，我才能夠任性地做我自己。



摘要

創造性任務因沒有正確答案且缺乏標準定義，讓許多人感到非常棘手，通常需要投入大量時間與精力去學習並精進專業知識與技能，才能夠完成此種任務。而且，在解決問題的過程中，必須取得外部回饋，並且透過不斷的迭代修改才能夠獲取高品質的結果。許多研究者已經能夠利用網路系統串連提供者與使用者，取得即時的外部回饋，幫助使用者完成任務。然而，大部分的研究只專注於回饋內容的改進，而忽略了最重要的目的，在於如何在此過程中輔助使用者學習，並有效的完成任務。所以，目的包含了獲得高品質的結果以及提高學習成效。此博士論文提出了一個解決創造性任務的迭代循環回饋的框架，分別探討產生有效且高品質回饋的生成機制以及有效整合回饋輔助編輯的方法，來促進有效的學習行為以及創造良好的使用經驗。我們設計與開發了一系列的智慧型回饋系統，運用群眾與機器合作的力量來幫助使用者快速取得符合需求的高品質寫作建議，並透過結構化的設計，引導他們有效地進行編輯，同時精進專業能力與提升作品品質。未來，在我們提出的互動回饋框架中，機器將擔任協調者與合作者的角色，根據使用者的偏好與行為，提供適當的回饋與引導，促進人與人以及人與機器良好的互動，一起合作完成困難的創造性任務。

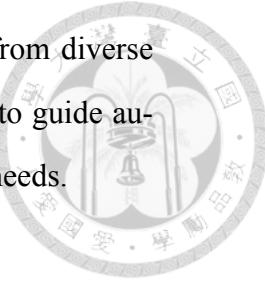




Abstract

Performing creative tasks is challenging, for such tasks are typically open-ended and ill-defined. To solve these complex problems, people need to spend much time and effort to learn professional skills and improve the in-progress work through an iterative process. Feedback is a critical component of this process for helping people discover errors and iterate toward better solutions. To meet the demand of timely feedback, recent work has explored technologies to connect problem solvers with feedback providers online. However, most research focuses on improving the content of feedback, but neglects the most important aspect of how to support problem solvers to learn and effectively facilitate the creation of high-quality outcome. In this dissertation, we explore several ways to support not only feedback generation but also feedback integration process, focusing on the writing tasks. We designed and developed intelligent systems that leverage the power of crowd and machines to support writers for obtaining effective feedback and facilitating good revision behaviors in the process. First, we start with our crowd-powered feedback system, *StructFeed*, and demonstrate a crowdsourcing approach for generating useful feedback to help writers resolve high-level writing issues in their revisions. Next, we propose feedback orchestration, which guides writers to resolve writing issues in a particular workflow by orchestrating feedback presentation. In the future, we would create an iterative feedback framework that enables collaboration between authors and feedback providers. Authors and feedback providers could benefit from this framework and collab-

oratively accomplish the complex creative tasks. The system would be regarded as a mediator for generating an ensemble of feedback from diverse feedback providers and for orchestrating feedback presentation to guide authors to achieve high revision performance based on individual needs.





Contents

口試委員會審定書	iii
誌謝	v
摘要	vii
Abstract	ix
1 Introduction	1
1.1 Background and Motivation	1
1.2 Supporting Creative Task Solving	2
1.3 Roadmap of Research	4
1.3.1 StructFeed: Generate Effective Feedback by Crowdsourcing	4
1.3.2 Feedback Orchestration: Facilitate Effective Revision	5
1.3.3 Understand How Feedback Affects Revision Results	6
1.3.4 Reflection Before/After Practice	7
2 Related Work	9
2.1 Human Computation and Crowdsourcing	9
2.1.1 Incentive, Microtasks, Quality Control	9
2.1.2 Crowdsourcing for complex task	11
2.2 Online Feedback Exchange	11
2.3 Learning Science	13
2.3.1 Self-Regulated Learning	13

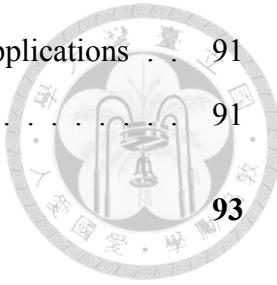
2.3.2	Reflection and Reflective Practice	14
2.4	Writing Support Systems	14
2.4.1	Automated Writing Evaluation	15
2.4.2	Crowd-Powered Systems for Writing Support	15
3	StructFeed: Generating Structural Feedback by Crowdsourcing	17
3.1	Introduction	17
3.2	StructFeed	18
3.2.1	Paragraph Unity and Topic Sentence	19
3.2.2	Crowdsourcing Workflow	20
3.2.3	Structural Feedback and Interface	22
3.2.4	Implementation	23
3.3	Unity Identification	23
3.3.1	Crowd-Based Method	24
3.3.2	ML-Based Methods	25
3.3.3	Evaluation	27
3.4	Field Deployment Study	29
3.4.1	Study Design	29
3.4.2	Tasks and Procedure	30
3.4.3	Measure	30
3.4.4	Results	31
3.5	Discussion	32
3.5.1	Crowd helps develop better rules for machine	32
3.5.2	StructFeed not only identifies writing issues but promotes reflection	33
3.5.3	Expert feedback performed worse than crowd feedback?	33
3.6	Conclusion	34
4	Feedback Orchestration: Supporting Reflection and Awareness in Revision	35
4.1	Introduction	35
4.2	Formative Study	38

4.2.1	Task and procedure	38
4.2.2	Findings	39
4.3	Feedback Orchestration	40
4.3.1	Expert revision practice	40
4.3.2	Support reflection and awareness	41
4.4	System Implementation	41
4.4.1	Taxonomy of writing feedback	42
4.4.2	Feedback classification	44
4.4.3	Revision workflow	45
4.4.4	Revision interface	45
4.5	Experiment	48
4.5.1	Participants	48
4.5.2	Task and procedure	49
4.5.3	Measures	51
4.5.4	Results	52
4.5.5	Insights from interviews with participants	53
4.6	Discussion and Implications	58
4.6.1	Feedback categorization guides learning behaviors	58
4.6.2	Flexible support for varying preference	59
4.6.3	Low-to-high sequence facilitates difficult problem solving	60
4.6.4	Decrease the workload of low-level, repetitive tasks	61
4.6.5	Resolving editing conflicts and mental obstacles	61
4.7	Limitations and Future work	62
4.8	Conclusion	63
5	How Feedback Affects Revision Quality?	65
5.1	Introduction	65
5.2	Crowd Feedback vs Expert Feedback	66
5.3	Experiment: Writing Revision Affected by Feedback of Different Types	67
5.3.1	Participants	69

5.3.2	Task and Procedure	69
5.3.3	Measures	70
5.3.4	Statistical Results	72
5.3.5	Insights from Interview Data	73
5.4	Discussion	75
5.4.1	The Cost and Benefit of StructFeed	75
5.4.2	The Utility of StructFeed Depends on Learners' Level of Proficiency	75
5.4.3	Gap between Expert Reviewer and Novice Writer	76
5.4.4	Macro-Task vs. Micro-Task	76
6	Reflection After/Before Practice	79
6.1	Introduction	79
6.2	Related Work	81
6.3	Learnersourcing for Drawing Support	81
6.4	ShareSketch: Draw, Review and Share	82
6.4.1	Sketch Interface	83
6.4.2	Timeline Interface for Sketch History	83
6.5	Before/After-Practice Reflection Workflow	83
6.6	Pilot Study	85
6.7	Results and Findings	85
6.7.1	After-practice annotation augments before-practice annotation . .	86
6.7.2	Before-practice reflection vs After-practice reflection	87
6.8	Discussion	87
6.8.1	Provide scaffolding for reflection and practice	87
6.8.2	Learning points as feedback enhances creative task learning . .	88
6.9	Future Work	88
7	Conclusion	89
7.1	Restatement of Contributions	89
7.2	Future Directions	90

7.2.1	Hybrid Combination of Crowd and Machine	90
7.2.2	Creative Knowledge Construction for Innovative Applications	91
7.3	Summary	91

Bibliography







List of Figures

1.1	Iterative feedback framework is designed for supporting creative task solving. It enables collaboration between authors and feedback providers.	2
1.2	The overview of this dissertation.	3
1.3	The concept of StructFeed.	4
1.4	The concept of feedback orchestration.	5
1.5	A series of studies to understand the relationship between revision quality and type of feedback generation method.	6
1.6	Reflection-based workflow is used to support learners to reflect and extract learning points from others' creation process.	7
2.1	A successful online feedback exchange framework requires five key elements: writing, feedback, revision, reflection as well as awareness. Our work incorporates reflection and awareness with a feedback framework for facilitating the effectiveness of revision.	13
3.1	The overview of StructFeed. The system generated writing suggestions based on aggregated crowd annotations and writing criteria.	19
3.2	The crowdsourcing interface contains 1) definition of topic sentence, 2) a worked-out example, 3) working area, 4) next button, 5) check-empty button, and 6) submit button.	21
3.3	The feedback interface contains 1) issue summary, 2) writing hints, and 3) topic and relevance sliders. The top image shows feedback when topic weight is 2 and relevance weight is 2; the bottom image shows feedback when topic weight is 3 and relevant is 0.	22

3.4	Results of topic sentence prediction by aggregating topic annotations from crowd workers with different threshold.	24
3.5	Results of relevant sentence prediction by aggregating relevant keyword annotations with different threshold.	24
3.6	Results of identifying irrelevant sentence by aggregating relevant keyword annotations with different threshold.	25
3.7	Results of topic sentence identification from ML-methods and our crowd-based method	28
3.8	Results of irrelevant sentence identification from ML-methods and our crowd-based method	29
3.9	The performance of three feedback-generation process is shown at the left three columns. The result of revision quality by three feedback-generation mechanisms is shown at the right three columns.	31
4.1	The overview of feedback orchestration. It guides writers to integrate feedback into revisions by a categorized structure.	37
4.2	The taxonomy of writing feedback follows the ESL Composition Profile. It consists of five writing evaluation criteria including content, organization, vocabulary, language use and mechanics. We grouped content and organization as high-level feedback, vocabulary and language use as medium-level, and mechanics as low-level feedback.	43
4.3	The two images shows the high-to-low interface that transits from the high-level stage (left) to the medium-level one (right). (a) original article view, (b) editing area, (c) feedback rating, (d) accept and reject buttons, (e) writing tip button, (f) next button, and (g) selected annotation.	46
4.4	Writers can find helpful information about writing by clicking the top-right button “writing tip.”	47

4.5 One example of expert feedback collected from one expert by using Word-service TOEFL writing feedback service. The feedback include one score, overall feedback and specific feedback containing direct editing and comments.	50
5.1 The overall comparison between an expert and an individual crowd worker.	67
5.2 The detailed comparison of feedback type between expert and crowd (individual view and overall view).	68
5.3 The detailed comparison of feedback form between expert and crowd (individual view and overall view).	68
5.4 Relation between Difference of Rating, English Proficiency Level and Feedback condition. Points shown are results predicted by the linear mixed-model but not the raw data. Fitting lines are added to illustrate the trends.	71
6.1 The overall of learnersourcing drawing support.	81
6.2 ShareSketch augments web-based drawing system with an interactive timeline. A user can create a drawing by a sketch interface (a), review the drawing process by an interactive timeline interface (b), and share the process to others (c).	82
6.3 A learner performs a reflection task by identifying a clip and describe what it is, and then explain why they learned from the clip.	84
6.4 Learners are allowed to practice drawing based on other's creation process in a short practice task.	84
6.5 The above two images are two creative processes created by two experts. The below eight images are the results of participants' practices.	86





List of Tables

4.1	Survey items for perceived helpfulness of overall system and feedback categorization.	52
4.2	The means (and standard errors) of perceived usefulness and helpfulness, as well as objective measures in each condition. There is no significant relationship between three conditions.	52
4.3	Non-novice writers preferred a workflow that is consisted to their previous revision strategies; novice writers changed their revision strategies after experiencing three types of workflows. The star symbol (*) indicates a novice writer who got low-level score and had less than one year of writing course experience.	56





Chapter 1

Introduction

1.1 Background and Motivation

Solving creative problems like writing or design is challenging, for such problems are open-ended and usually may not have well-defined answers. The answer is not true or false, but how good the answer is. Therefore, evaluating the answer is also a difficult problem. The quality of answer usually can be evaluated by multiple criteria, yet the criteria may be change under distinct situations. For example, writing an essay or creating a resume requires very different criteria to evaluate the achievement. To achieve high-quality results, people have to acquire domain-specific knowledge and develop professional skills by learning and practicing repeatedly. As problem gets large and complicate, people often need external support to accomplish such complex creative tasks.

Feedback is a critical component in the creative process for supporting people iterate toward better results. It helps people acquire conceptual knowledge, discover weakness, and fix errors. However, obtaining high-quality feedback is difficult due to a limited pool of experts. To meet the demand of obtaining timely feedback, recent work has explored online feedback exchange platforms to connect problem solvers with feedback providers from diverse sources [72, 50, 27].

Crowdsourcing has been applied and proven to be useful for helping designers improve their design by diverse feedback obtained from the crowd [72, 51, 50]. Most studies focus on how to generate multiple types of useful feedback including high-level impres-

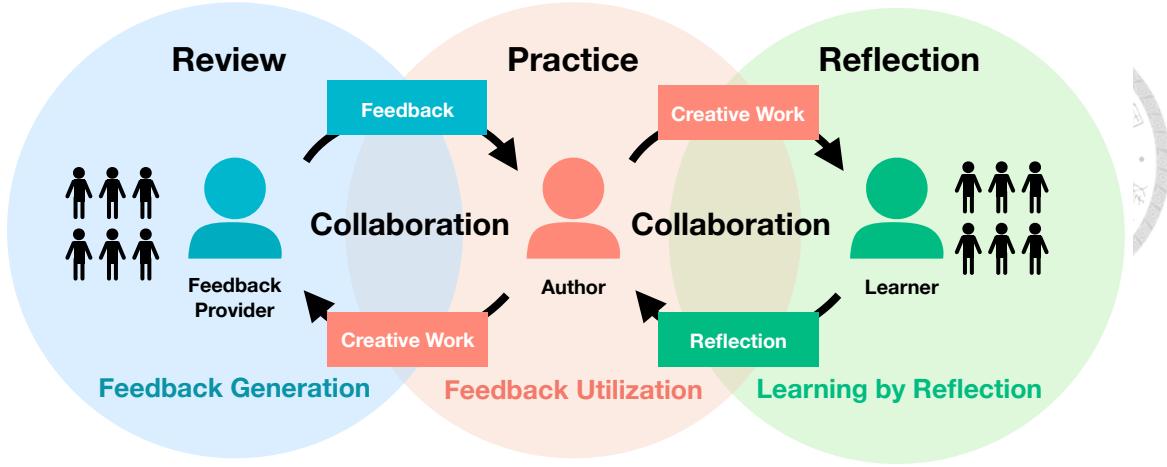


Figure 1.1: Iterative feedback framework is designed for supporting creative task solving. It enables collaboration between authors and feedback providers.

sions and concrete suggestions regarding design principles. In addition, crowdsourcing workflows and structural interfaces were proposed for guiding crowd workers to generate high-quality results. These studies have demonstrated that designers have benefited from diverse feedback generated by the crowd in the design process. However, diverse feedback not always leads to good results. Most recent studies focus on feedback generation process, but ignore feedback integration process. Therefore, we attempts to provide a complete understanding of online feedback exchange and explore possibilities of crowd-powered feedback systems to from both aspects.

In my dissertation, I build on studies of learning science and develops intelligent systems that generate effective feedback and facilitate good revision behaviors, focusing on the writing domain. The goal is to leverage the power of crowd and machine to support people to solve the complex creative tasks from diverse aspects and improve the task quality effectively; more importantly, the work aims to provide people better learning experiences in the iterative problem-solving process.

1.2 Supporting Creative Task Solving

To solve creative tasks, we propose an iterative feedback framework that enables collaborations between authors and feedback providers. In this framework, authors can learn to

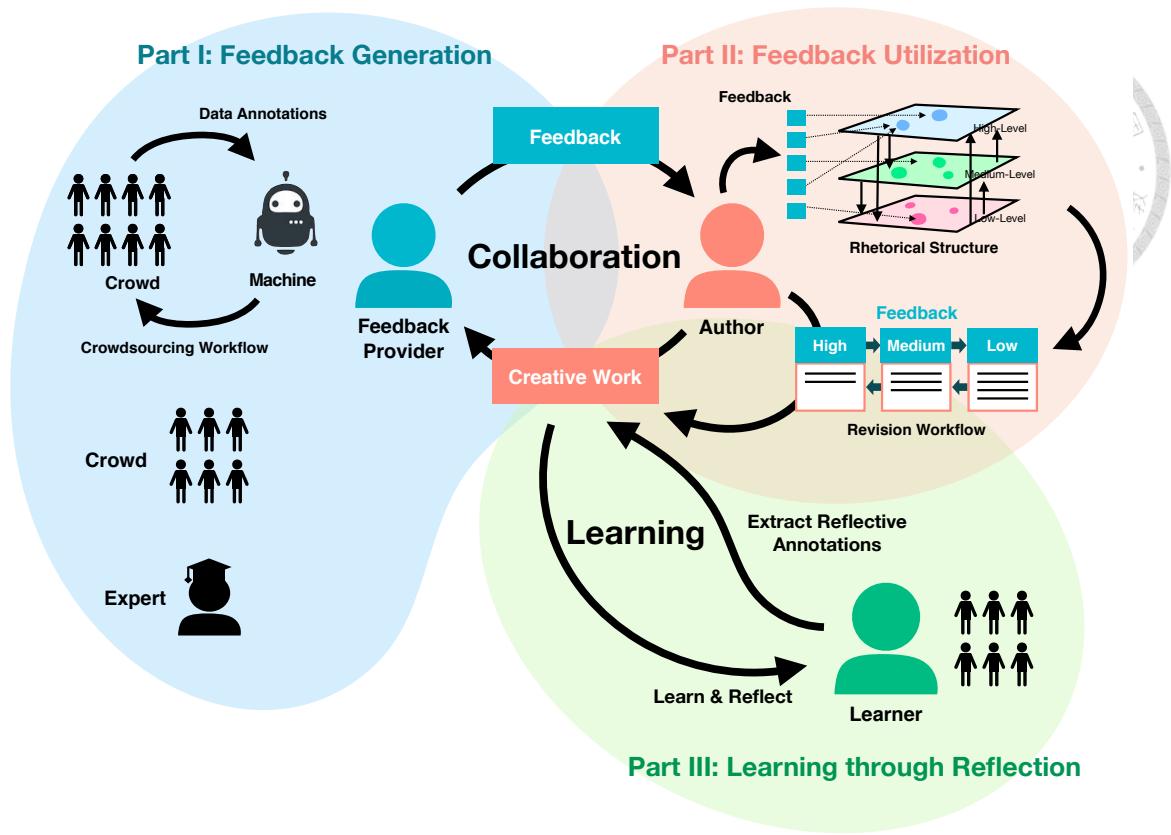


Figure 1.2: The overview of this dissertation.

improve the quality results by feedback obtained from the feedback providers; feedback providers can also learn to evaluate the quality of outcome and provide helpful feedback to authors. The overview of the framework is depicted in Figure 1.1. Both roles of people can learn from each other and collaborate to achieve the best performance in the iterative process.

In my dissertation, I structure this problem into four aspects: 1) understand how to generate effective feedback for supporting creative tasks solving, and 2) investigate how to effectively integrate feedback and facilitate high-quality results, 3) understand how different types of feedback affects the revision quality in the iterative process, and 4) explore how to promote reflection to enhance learning process.

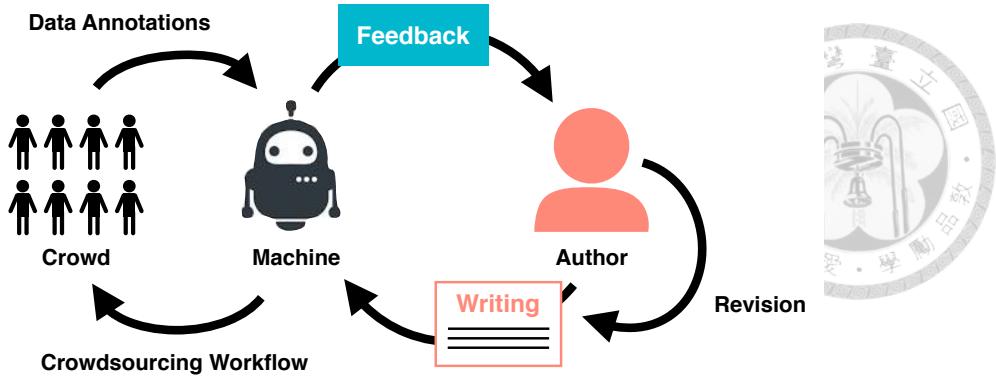


Figure 1.3: The concept of StructFeed.

1.3 Roadmap of Research

The section presents a brief overview of this dissertation (see Figure 1.2). First, we focus on feedback generation process and introduce *StructFeed*, a crowd-powered system that generates effective feedback for writing support. Second, we explore feedback integration process and propose *feedback orchestration* to facilitate good revision and learning behaviors. In the end, we will explore a series of studies to understand the significant factors that impact the creative task performance in the iterative process. This dissertation will contribute rich insights of designing a successful intelligent feedback systems for supporting creative tasks.

1.3.1 StructFeed: Generate Effective Feedback by Crowdsourcing

StructFeed is a crowd-powered system that supports non-native writers to create a unified articles by providing useful suggestions including writing hints and crowd annotations. In this project, we consulted with English writing teachers and textbooks to find the key principles of writing evaluation criteria. Based on the understandings, we designed a crowdsourcing workflow with a goal of evaluating unity of writing, which is the most important criterion of good writing. The workflow decomposed the evaluation tasks into micro-tasks and allowed multiple crowd workers to annotate topic sentence and relevant keywords. Next, the system identified topic and irrelevant sentences based on aggregated results. Last, it generated suggestions based on writing criteria. The concept of StructFeed

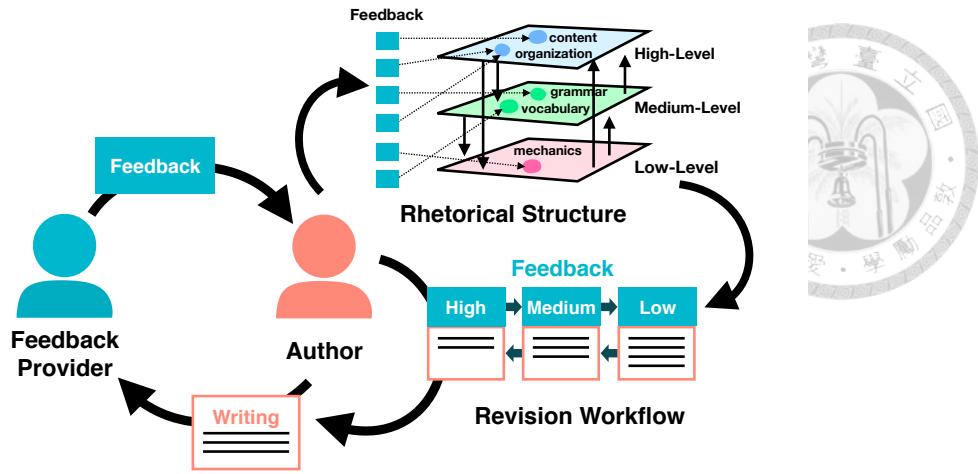


Figure 1.4: The concept of feedback orchestration.

is shown in Figure 1.3.

In a field experiment, we recruited 18 self-motivated non-native speakers to write an independent essay and revise the essay by feedback obtained from experts, crowds, and StructFeed. We compared the quality of the original and revised article in three groups. The results showed that non-native writers who received feedback from StructFeed got the best performance, compared with feedback from a single expert or crowd worker. Two interesting findings are as follows:

- Feedback provided by experts did not lead to the best revision results, despite the fact that it got the highest scores of perceived usefulness.
- Providing indirect feedback containing writing hints about subgoal and aggregated crowd annotations, which is generated by StructFeed, promotes people to reflect their deficiency, resulting in better revision results.

1.3.2 Feedback Orchestration: Facilitate Effective Revision

Diverse feedback provides useful information that helps people improve task performance. However, good feedback may not lead to high-quality outcome. In this work, we attempt to understand people's behaviors in the feedback integration process and explore ways to facilitate effective revision behaviors by adjusting feedback presentation. We proposed

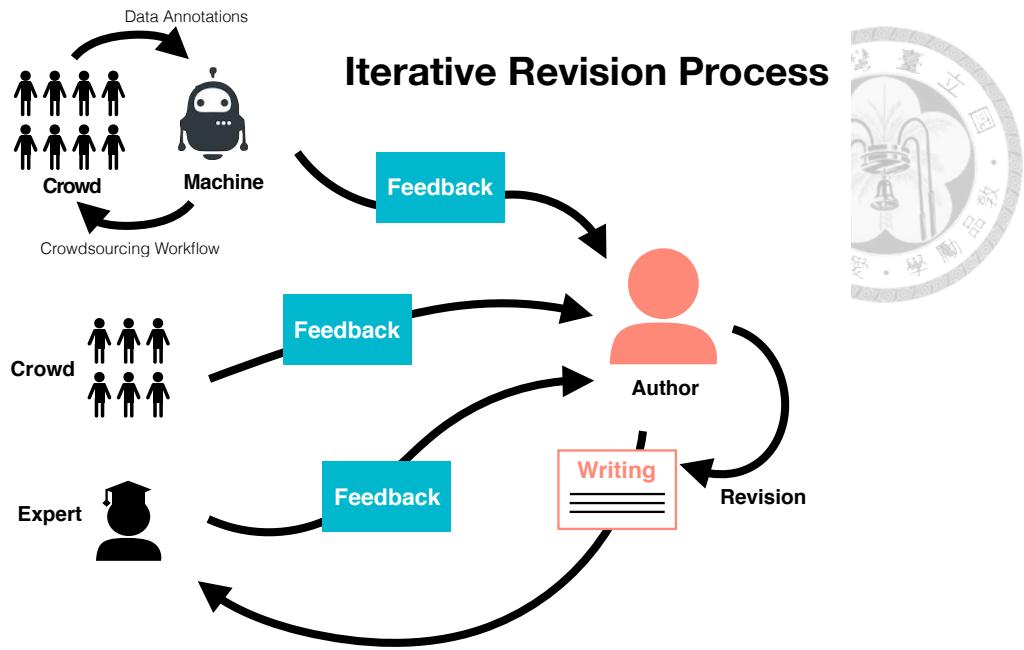


Figure 1.5: A series of studies to understand the relationship between revision quality and type of feedback generation method.

feedback orchestration that uses a categorized structure to guide writers to learn and integrate feedback in a revision workflow (see Figure 1.4). To evaluate the idea, we implement ReviseO, a web-based writing support system that use a standard writing rubric to classify feedback and enable sequential and concurrent revision workflows.

In a within-subjects experiment, we explored how three types of feedback presentation (high-to-low, low-to-high, and all) affect revision behaviors and task performance. The results showed that our system guided people to improve their writings and helped people discover their weaknesses of writing. Furthermore, The findings suggested that people receiving feedback separately spent more time but fewer edits on their revision, compared with people receiving all types of feedback together. In addition, people receiving feedback in the low-to-high process developed momentum to solve the harder problems like content and organization.

1.3.3 Understand How Feedback Affects Revision Results

According to our previous research, we observed that expert feedback performed worse than crowd feedback when all participants had high satisfaction about obtaining it. In this

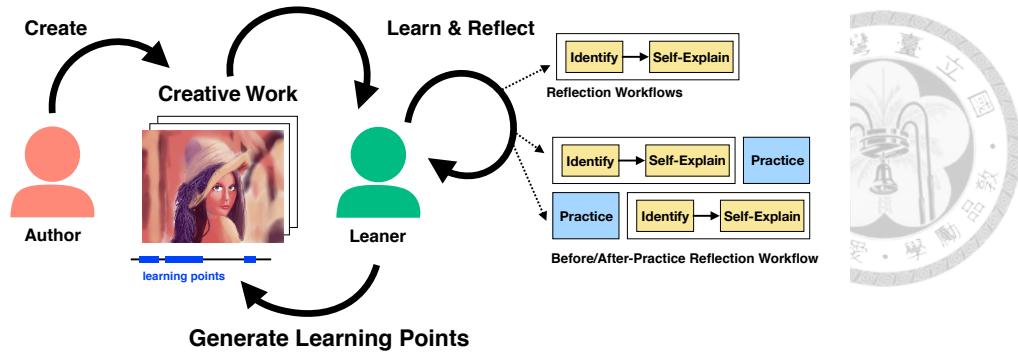


Figure 1.6: Reflection-based workflow is used to support learners to reflect and extract learning points from others' creation process.

project, we are interested in how difference of expert feedback and crowd feedback, and which one can benefit the writers in what ways. which factors significantly influence the quality of results.

To better understand how feedback generated from different sources affects revision quality (see Figure 1.5). First. we conducted a content analysis on two groups of feedback collect from experts and crowd workers. We observed that although a single expert can generate more types of feedback than a single crowd workers, both of them tend to less high-level feedback about content and organization. Next, we will design a within-subject experiment to examine the relationship between type of feedback, order of feedback type, and writers' English abilities.

1.3.4 Reflection Before/After Practice

To support the acquisition of drawing skills, this research explores a learnersourcing approach to generating personalized learning points. These are annotations containing a clip of a drawing process, a description, and an explanation. This paper presents ShareSketch, a web-based drawing system that enables learners to practice drawing, review the drawing process, and share their works with others. In particular, we propose the before/after-practice reflection workflow that allows learners to generate learning points before or after each short practice. We evaluated our reflection workflow with eight self-motivated drawing learners. The results showed that our reflection workflow can guide learners to

generate high-level subgoal or concept labels, low-level steps, and personalized coping strategies.





Chapter 2

Related Work

This dissertation builds upon four main research areas: **human computation and crowdsourcing** to incorporate human intelligence to solve the complex creative tasks in a distributed way; **online feedback exchange** to solicit feedback for improving the quality of creative tasks; **learning science** to make feedback and revision effective for supporting novices to learn domain knowledge, develop specific skills, and generate high-quality content; **writing support applications** to better understand the challenge of designing intelligent systems for supporting creative tasks.

2.1 Human Computation and Crowdsourcing

Over the last decades, human computation has established a popular research area that harnesses human intelligence to solve difficult problems that are beyond the scope of existing Artificial Intelligence (AI) approaches. Many researchers have successfully leveraged human abilities to tackle the problems that computers cannot solve, from image labeling [70], to protein folding [11], to text digitization, and to word processing [1].

2.1.1 Incentive, Microtasks, Quality Control

Most human computation systems solicit human inputs for a part of computation in a distributed way. They treat humans as processors, but unlike machine, humans require an incentive to make contributions to a computation goal. As a result, one of challenges for

designing human computation systems is incentive design. The incentive can be various forms including monetary or social rewards, enjoyment of task, a sense of duty or a desire of learning. For example, Game with a Purpose (GWAP) attracts online players to perform some specific tasks (e.g., tagging object from an image [70] or protein folding [11]) for entertainment as playing a game. In addition, many human computation systems may recruit paid workers through an online labor market such as Amazon Mechanical Turk or CrowdFlower, in which workers earn small amount of money by accomplishing “micro-task” in a short time. Another systems allows friends or followers on social media sites such as Facebook or Twitter to make contributions based on social relationship [2]. Furthermore, volunteering-based systems motivate non-paid crowds to make contributions for the goal that is serving the public good. For example, Zooniverse¹ engages volunteers to help professional reserchers to perform various science tasks including classifying galaxies from photos, identifying crater entries, and so on. Learnersourcing engages a large group of leaners to contribute their efforts for improving video contents and interfaces in the context of learning [71].

Instead of incentive, human attention and task complexity are another important design considerations. Attracting a large number of crowd workers to perform an arbitrary task may lead to noise data, because of limited attention and a lack of expertise. To handle noise data, human computation systems aim to apply “divide and conquer” strategy to break large problems into smaller subtasks, called microtask, and devise various quality control mechanisms for improving the quality of outcome from large amounts of crowd workers.

While human computation systems focus on leveraging human intelligence to solve the task that machine cannot solve, crowdsourcing systems addresses this issues in a scalable way by recruiting a group of crowd workers. This work is inspired by previous research about crowdsourcing and aims to design crowd-powered systems that support novices to solve complex creative tasks and develop their professional skills.

¹<https://www.zooniverse.org>

2.1.2 Crowdsourcing for complex task

To allow online crowd workers to solve complex task, existing crowd-powered systems break complex into subtasks and guide workers to perform the subtasks in a workflow.

For example, Soylent [1] has demonstrated that dividing complex tasks into various small tasks in a multi-stage process improves the quality of document editing. CrowdForge [41] uses a MapReduce framework to guide workers to write encyclopedia articles. Other workflows are used to retrieve nutrition information from food photographs [56], taxonomy generation [10], and travel planning [76]. These workflows can even be created by workers themselves [44]. However, some tasks cannot be divided into independent tasks and may lose global context or information while applying such decomposition strategies.

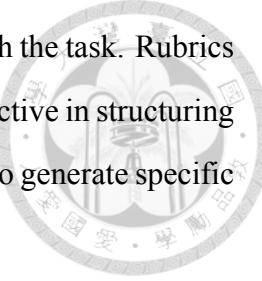
To addresses these issues, many researchers explored ways to enable crowd collaboration to solve complex tasks requiring global context [38, 69, 29], such as travel planning, short story [38] or wikipedia article writing [29], etc. Ensemble [38] used a team leader to coordinate a group of crowd workers to write short stories. The Knowledge Accelerator [29] developed several crowdsourcing techniques like “vote-then-edit” and “task of least resistance” to guide crowd workers to enhance global consistency of writing. Flash teams [61] or Turkomatic [44] have relied on a invested contributor such as a moderator or an experienced contributor to maintain the big picture.

This thesis contributes crowdsourced workflow design for allowing crowd workers to examine quality of creative work and generate annotations for effective writing feedback.

2.2 Online Feedback Exchange

Feedback is critical to the success of creative work, from essay writing to design [63, 45]. The CSCW community have explored Online Feedback Exchange (OFE) systems for requesting and receiving information from distributed feedback providers to help designers improve the quality of creative work [24]. However, the feedback collected from different sources, including online communities, crowds and peers, can be general and ambiguous because those providers may have different motivations, expertise and perspectives [73].

To collect timely, specific and useful feedback, prior work used multiple-stage workflows [72] or a structured interface [50] to guide novice workers to accomplish the task. Rubrics [50, 27, 75] and comparative examples [43] have been proven to be effective in structuring feedback generation process because they encourage novice providers to generate specific suggestions based on diverse criteria.



Furthermore, many researchers have investigated the characteristics of feedback that can affect feedback reception and work performance [54, 43, 32, 30]. In addition to specificity, feedback framed with positive language from an anonymous source improves perceived helpfulness of design critique [54, 32]. The motivation of providers also affects the quantity, quality, and content of feedback. For example, a paid task market provides longer and more positive suggestions and a web forum provides more process-oriented feedback [73]. Moreover, researchers have developed a natural language model based on these characteristics (e.g., specificity, sentiment, etc.) to provide guidance with feedback examples to help providers self-assess and improve their critique [43].

Although these studies have improved the perceived helpfulness of feedback, some designers still failed to improve their design work [72, 50]. Researchers found that novice designers tended to focus on solving minor, easily-identified issues rather than important ones [50]. Therefore, recent work have started to investigate the relationship between feedback provided and the actual performance, especially focusing on understanding the effectiveness of different approaches to interpreting feedback [25, 74]. Foong et al. have found that expert and novice designers make sense of feedback in different ways and suggest different forms of support for them to successfully integrate feedback into revision [25]. In addition, Yen et al. have shown that combining feedback review with a reflective activity can support deeper interpretation of feedback [74].

Our research also investigates how feedback receivers interpret and integrate feedback in revision; however, we focus on understanding writers' revision behaviors and developing approaches to facilitating reflection and awareness, leading writers to take advantage of feedback and revise effectively. Figure 2.1 describes how our work relate to prior work. Our work regards reflection and awareness as key elements for designing a successful on-

line feedback exchange system because reflection and awareness can help enhance the effectiveness of revision.

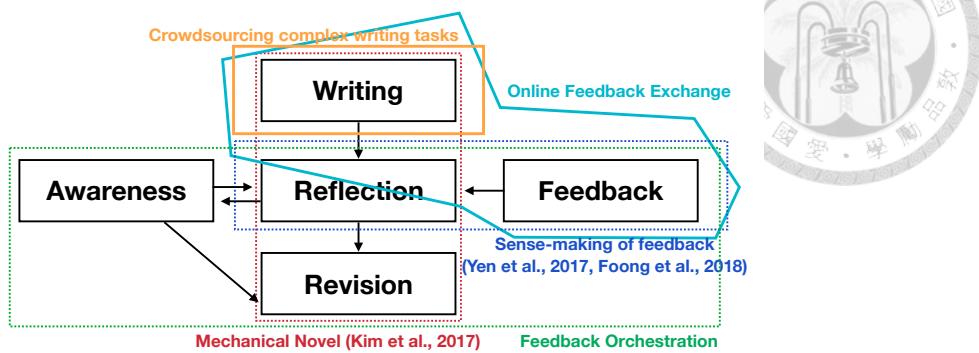


Figure 2.1: A successful online feedback exchange framework requires five key elements: writing, feedback, revision, reflection as well as awareness. Our work incorporates reflection and awareness with a feedback framework for facilitating the effectiveness of revision.

2.3 Learning Science

2.3.1 Self-Regulated Learning

To support successful learning on creative tasks, self-regulation should be considered in the process. This work draws on the idea of self-regulated learning. There is considerable research evidence suggesting that effective feedback leads to learning gains. In particular, Sadler identified three conditions for effective feedback. He argued that good feedback must help a student to 1) possess a concept of the standard being aimed for, 2) compare the actual level of performance with the standard, and 3) engage in appropriate action that closes the gap [63].

Self-regulated learning has been regarded as an important component for successful learning in school and beyond [4, 77]. Pintrich defined self-regulated learning as “an active, constructive process whereby learners set goals for their learning and then attempt to monitor, regulate, and control their cognitions, motivation, and behavior, guided and constrained by their goals and the contextual features in the environment” (p.453) [5]. To support learners to take control of their own learning, Nicol [55] identified seven principles of good feedback practice that support self-regulation: 1) clarify what good performance

is (goal/criteria/standards); 2) facilitate self-assessment; 3) deliver high-quality feedback; 4) encourage dialogue between teachers and peers; 5) encourage positive motivation and self-esteem; 6) provide opportunities to close the gap; and 7) use feedback to improve teaching.



2.3.2 Reflection and Reflective Practice

Reflection is defined as a purposeful thinking toward a goal [13]. John Dewey argued that reflecting on experience facilitates learning from experience. It is regarded as a mental activity of internal problem-solving. Furthermore, Donald A. Schön introduced reflective practice and explained the relationship between reflection and practice. He theorized that reflective practice represents an important factor to improve professional activity and distinguished two types of reflection: **reflection-in-action** and **reflection-on-action** [64, 65].

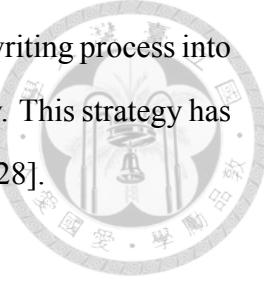
Reflection-in-action indicates that students monitor and modify actions during the learning process. For example, students are encouraged to think about their current learning performance and figure out achievable steps to accomplish the goal. On the other hand, **reflection-on-action** indicates that students evaluate the past action and make a plan to improve future actions after the learning process. For example, students are encouraged to think back to their previous learning process and identify what was done well and what could have been done better.

To facilitate effective revision, our systems apply a rhetorical structure, provide meta-feedback and structural feedback to help people be aware of their weaknesses, and reflect in or on actions to develop their strategies. In addition, this work incorporates reflection into the system design and presents a reflection workflow and a reflection-before/after-practice workflow to solicit annotations for supporting learning behaviors.

2.4 Writing Support Systems

Writing is considered as a series of cognitive processes, including planning, translating, reviewing and monitoring [22]. These processes were distinctive and hierarchically orga-

nized; however, any process can be embedded within any other, leading writers to create highly personal and complex workflows. Breaking down the complex writing process into smaller, more manageable subtasks is a commonly used writing strategy. This strategy has been shown to reduce cognitive load and improve writing quality [52, 28].



2.4.1 Automated Writing Evaluation

State-of-the-art automated writing evaluation systems (AWE) for supporting second language writing utilize supervised learning on large training datasets to predict the holistic score and to generate diagnostic feedback of an essay [14]. As an example, ETS Criterion [8] uses a discourse structure trained on 1,462 labeled essays, and builds a specific scoring model for each topic trained on a sample of 200–250 labeled essays. Consequently, it only supports limited topics for writing practices due to the cost of training data collection, annotation, and processing. The proposed crowd-based writing framework is scalable, which supports any topic by collecting annotations on key elements of an essay (e.g., topic sentences, keywords) and providing structural feedback in alignment with the principles of writing.

2.4.2 Crowd-Powered Systems for Writing Support

Many researchers have applied such strategies to divide writing into a series of micro-tasks and have allowed crowd workers to accomplish them using a workflow [48, 1, 41, 38, 39, 67, 53]. For example, Soylent [1] split writing into stages and guided crowd workers to shorten texts, proofread and provide writing suggestions. CrowdForge [41] used MapReduce-like framework to generate encyclopedia articles by guiding multiple independent workers to perform simple tasks such as creating an outline, collecting facts and writing paragraphs. Storia [39] used a narrative structure to guide novice crowd workers to write a short story based on social media content. In addition, WearWrite [53] demonstrated the feasibility of writing a paper using a smartwatch by using microtasks and online crowd workers. Nevertheless, creating coherent and comprehensive writing can be challenging by microtask design because crowd workers lack a global view.

To address this issue, researchers have explored several approaches that promote collaborative work and allow crowd workers to perform microtasks while achieving high-level goals [38, 69, 29, 40]. Ensemble [38] used a team leader to guide crowd workers to write short stories. The leader was used to direct the overall collaborative process by establishing creative goals. MicroWriter [67] focused on coordinating a group of collocated people to write a single report with a shared goal via microtasks. In addition, some crowdsourcing techniques were also introduced to enhance global consistency of writing. Voting activities can be used to help workers think about work from a global viewpoint and organize ideas to achieve high-level goals [29]. Guiding crowd workers to reflect on a high-level goal and revising the work by splitting it into low-level, actionable tasks can lead to high-quality results [40].

Those studies focus mainly on the challenges of breaking down the complex writing process into smaller, easier subtasks and aim for enabling collaboration between writers to accomplish the tasks. Although some researchers have noticed the effectiveness of incorporating reflection and revision in the writing process, novice writers still cannot fully benefit from these activities without any intervention [23]. Our research addresses the difficulties of revision and scaffolds novice writers to self-assess their work based on a rhetorical structure and structured feedback and support them to think and revise effectively.



Chapter 3

StructFeed: Generating Structural Feedback by Crowdsourcing

3.1 Introduction

Writing is a difficult task, especially for non-native speakers. To construct a well-structured essay, ESL writers usually take much efforts and time to write and rewrite iteratively. During the iterative process, many written compositions which have many writing issues like lack of clarity or focus, or incomplete topic development are generated by ESL learners. They need a pair of outside eyes to identify their weak spots and to suggest ways of fixing them. However, collecting high-quality feedback is challenging due to the limited pool of experts available. To support on-demand help, we require an approach that supports ESL writers to identify writing issues and suggest ways for improving their writing.

Previous studies have explored automated writing evaluation systems (AWE) for supporting second language writing. Almost all studies utilize supervised learning on large training datasets to predict the holistic score and to generate diagnostic feedback of an essay [14]. However, those automated methods only support limited topics for writing practices due to the cost of training data collection, annotation, and processing. To enable diverse writing support, we leverage the power of native speaker to make small contributions for identifying basic writing elements like topic sentences and supporting data in the

texts.

According to our observations, many ESL students with more than 10-year English learning experience still struggled with identifying topic sentence for a basic five-paragraph essay and usually failed to develop a unified essay in our pilot study. That is why the ESL writing pedagogy always starts with teaching topic sentence writing, and move on the development of paragraph and essay later [57]. Therefore, we propose StructFeed, a crowd-powered system that generates structural feedback for helping ESL writers recognize high-level writing issues and produce a unified article. A crowdsourcing workflow is used for allowing native-speakers to identify topic sentence and relevant keywords in an article. Next, the system will predict the location of topic sentences and irrelevant sentence by aggregating crowd annotations and then generate writing suggestions. The goal is to guide people to revise the paragraph to achieve the paragraph unity based on writing criteria.

We compare our crowd-based method with three naïve machine learning (ML) methods. The results suggest that the crowd-based method outperforms all ML methods. In addition, the new rule derived from crowd annotations outperformed all initial methods. Furthermore, we evaluate StructFeed with 18 ESL writers recruited through online postings in the community sites. A between-subject experiment was conducted to investigate how and whether people can improve their writing after receiving feedback generated by one crowd worker, one expert, or StructFeed. The results showed that people who received writing suggestions from StructFeed achieved the best performance than other people who received writing suggestions from one crowd worker or one expert.

3.2 StructFeed

StructFeed is a crowd-based system that allows a user to request, receive, and review writing feedback for recognizing and fixing structural issues of writing. Instead of providing feedback on local issues like grammatical or spelling errors, StructFeed attempts to address global issues like irrelevant ideas or missing main topic.

In this section, we introduce the design of StructFeed, and the overview of the system

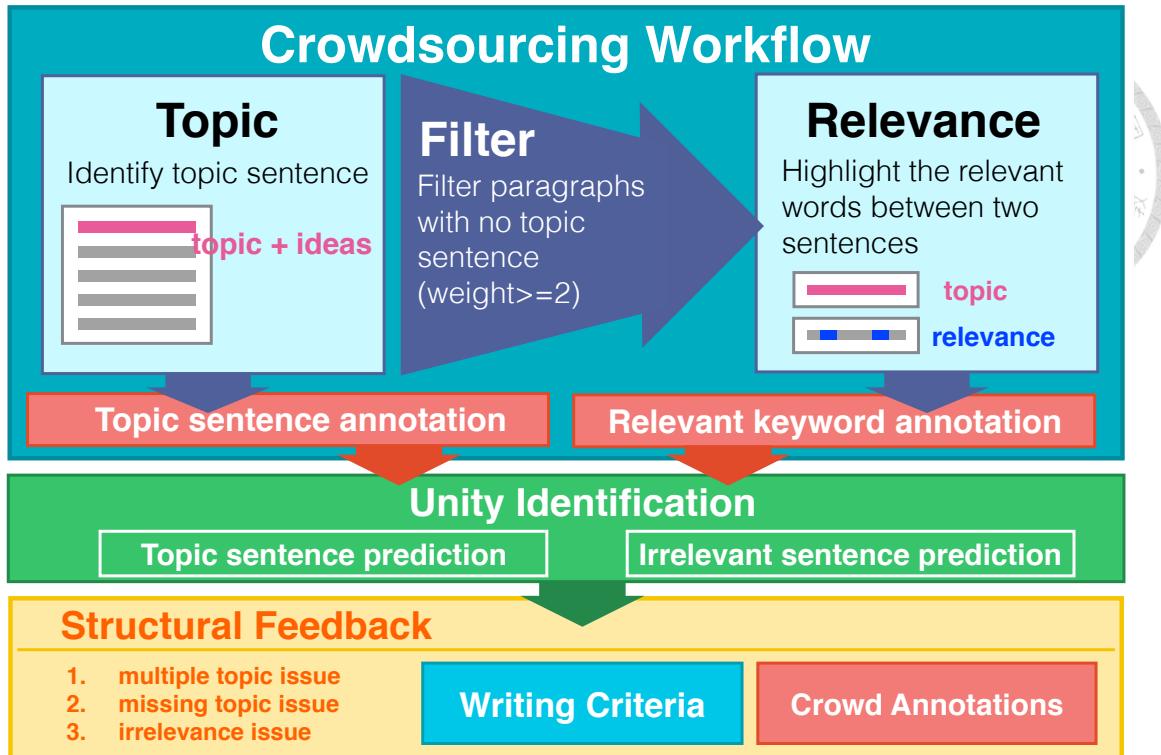


Figure 3.1: The overview of StructFeed. The system generated writing suggestions based on aggregated crowd annotations and writing criteria.

is depicted in Figure 3.1. First, we describe the essential principle of writing – paragraph unity. Next, we introduce our crowdsourcing workflow that allows crowd workers who are native speakers to examine the paragraph unity through two types of micro-tasks. Finally, we present the structural feedback with a visualization interface.

3.2.1 Paragraph Unity and Topic Sentence

A good essay should have a clear structure in which all elements are well organized and linked. An essay consists of introduction, body, and conclusion and each part is composed of paragraphs. A paragraph is the basic component of writing, and it is a group of related sentences that are organized to develop a single idea. It contains a topic sentence, several supporting sentences, and a concluding sentence. The topic sentence is the most important one because it indicates the main idea of a paragraph. The supporting sentences are used to provide evidence to support the main idea. The concluding sentence is used to summarize the main idea presented in the topic sentence and emphasize the impression on the readers.

A good paragraph should follow an important principle called unity. Unity is used to evaluate the quality of oneness in a paragraph or an essay. It can be achieved by the following two steps.

- All sub-points are related to one main idea.
- No irrelevant sentence exists in the paragraph.



3.2.2 Crowdsourcing Workflow

The designed workflow breaks down the process of unity identification into two stages: Topic and Relevance stage.

The system dispatches micro-tasks to online crowdsourcing marketplace in both Topic and Relevance stages. There is a filter between the two stages. It aggregates results from Topic stage and passes qualified results to Relevance stage.

Topic Stage

The goal of Topic stage is to examine whether all paragraphs have a topic sentence. In this stage, the system creates a task with five assignments and distribute it to distinct crowd workers. The task asks workers to mark every topic sentence in an essay. Our tool lets workers make sentence-level annotation by clicking on any part of a candidate sentence. The selected sentence will be highlighted with yellow background. The annotation can be cancelled by re-clicking.

The crowdsourcing interface in Figure 3.2 is designed to guide workers to accomplish the task with good quality. The interface contains a brief description of topic sentence (1), a worked-out example (2) for teaching workers how to identify topic sentence, and a working area (3). In the working area, a crowd worker can annotate a sentence with a simple click. A next button (4) is used to make workers focus one paragraph at a time; when it is clicked, the next paragraph will appear in the working area. When all the paragraphs appears, the check-empty (5) and submit button (6) will show up. In the end, a worker can submit the answer and finish the task.

Task: Identify Topic Sentence

1. What is a topic sentence?

- A topic sentence is a sentence that captures the meaning of the entire paragraph or group of sentences.
- It is composed of two parts: (a) the topic itself and (b) the controlling idea.

2. Example

Title: Affective Computational Priming and Creativity

Content:

Creative professionals, such as engineers, scientists, and designers, increasingly rely on technology to complete creative tasks at work. They use web browsers to conduct research and search for inspiration. They document projects on wikis in order to collaborate with others. They depend on graphic design software and word processors to design and create. In response, human-computer interaction (HCI) researchers have started building creativity support tools, or software to enhance creative performance [cited]

3. Please identify topic sentence in each paragraph.

If no topic sentence in the paragraph, please skip the paragraph. If no any topic sentence in the whole essay, please indicate "No topic sentence" by the checkbox at the bottom page.

Title: Has human harmed the Earth?

Content:

Although the quality of life has improved over the past decades due to new technological advances but the damages made to the earth weigh more. Damages include increase in pollution and change in climatic patterns. Irreversible damage to earth can include depletion of natural resources.

As the technology advances more factories are built. These factories dispose waste material into natural water, which could be harmful to aquatic life. Emissions from the factories and automobiles pollute the air, which we breathe. Nuclear waste and radiation from power plants are harmful to our health.

4. Go to next paragraph

5. No topic sentence in the essay

6. Submit

Figure 3.2: The crowdsourcing interface contains 1) definition of topic sentence, 2) a worked-out example, 3) working area, 4) next button, 5) check-empty button, and 6) submit button.

Relevance Stage

The goal of Relevance stage is to determine whether every other sentence is related to the topic sentence in a paragraph.

In this stage, the system creates a task with three assignments and dispatch it to different workers. The task contains one paragraph with topic sentence labeled. The given topic sentence is determined by majority voting from the previous stage and is highlighted in yellow color. The task asks workers to locate the word which is related to the given topic sentence. Similar to the design of the previous stage, workers can make word-level annotation by clicking on any part of a candidate word. The selected sentence will be highlighted with a green background. The annotation can be canceled by re-clicking.

Filter

Filter is a bridge component which aggregates all annotations generated from the Topic stage and determines which one is a topic sentence by at least two annotations labeled from

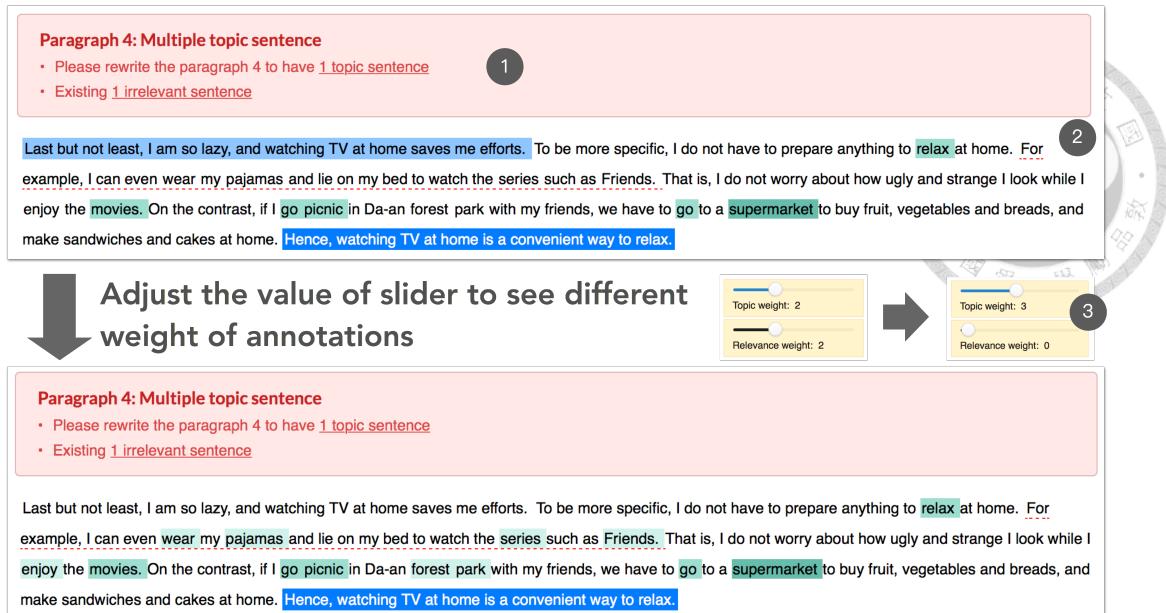


Figure 3.3: The feedback interface contains 1) issue summary, 2) writing hints, and 3) topic and relevance sliders. The top image shows feedback when topic weight is 2 and relevance weight is 2; the bottom image shows feedback when topic weight is 3 and relevant is 0.

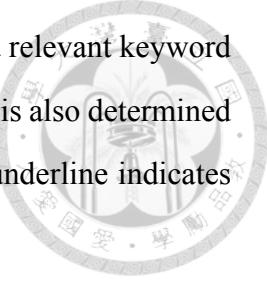
different workers. Next, the Filter would choose paragraphs existing a topic sentence to pass them to the Relevance stage.

3.2.3 Structural Feedback and Interface

Structural feedback is designed for helping writers identify their writing issues and facilitate rewriting behaviors by prompting writing hints. The feedback consists of two elements: issue summary (1) and writing hints (2). The issue summary indicates the type of writing issue including *multiple topics issue*, *irrelevance issue*, and *missing topic issue*, and a suggested editing action (see Figure 3.3); the writing hints show the detailed of writing issues by a number of low-level annotations. The annotations include topic sentences, irrelevant sentences, and relevant keywords. The design of writing feedback follows Sadler's requirements for high-quality feedback [63].

We not only show the location but also the weight for the annotations of topic sentences and relevant keywords. The weight of annotation presents the number of agreements made from different people. The blue highlighted sentence is topic sentence. The brightness of background color indicates the number of agreement from different people. When

more people annotate the same sentence as topic sentence, the background color of this sentence is much deeper than the other. In addition, the annotation of a relevant keyword is indicated by green highlighting. The brightness of background color is also determined by the number of annotations generated by workers. The red dotted underline indicates the location of an irrelevant sentence.



The two sliders (3) at the top left corner of the page are used to filter two types of annotation by different weight. By moving the slider back and forth, the writer can see the annotations with different weight appears in sequential order (see Figure 3.3).

3.2.4 Implementation

StructFeed is a Web application built in Python, Javascript, and Postgres, which has been deployed on Heroku. The two types of micro-tasks in the workflow generated as two external HITs are submitted to Amazon.com’s Mechanical Turk, a popular online crowdsourcing platform. Workers who have at least 80% task approval rate are considered to perform our tasks. Each task costs \$0.05 and one worker can perform 2.5-3 tasks in a minute. The worker can get at least \$7.5-\$9 per hour (higher than \$7.25).

3.3 Unity Identification

To support writing on any topic, StructFeed needs to work in the absence of enough data, i.e. the “cold-start problem” in the context of computer-assisted writing. Therefore, we compare our crowdsourcing approach of unity identification with three naïve machine learning (ML) methods. These ML methods are commonly used for solving problems without large amounts of labeled training data.

In this section, we evaluate all methods by calculating average precision, recall, and F1-score (combined precision and recall) of identifying topic sentence, and irrelevant sentence.

Agreement	Precision	Recall	F1 score
1	0.459259	0.826667	0.590476
2	0.606742	0.720000	0.658537
3	0.655738	0.533333	0.588235
4	0.774194	0.320000	0.452830
5	0.733333	0.146667	0.244444



Figure 3.4: Results of topic sentence prediction by aggregating topic annotations from crowd workers with different threshold.

Agreement	Precision	Recall	F1 score
1	0.711628	0.884393	0.788660
2	0.718085	0.780347	0.747922
3	0.722581	0.647399	0.682927
4	0.698925	0.375723	0.488722
5	0.772727	0.196532	0.313364

Figure 3.5: Results of relevant sentence prediction by aggregating relevant keyword annotations with different threshold.

3.3.1 Crowd-Based Method

Topic/irrelevant sentence prediction

A topic sentence is directly determined by at least two distinct crowd workers. Relevant sentence is determined by at least three distinct crowd workers who clicked the same sentence including relevant keywords. The threshold of a topic or relevant sentence is determined by empirical data discussed in the following paragraph. An irrelevant sentence is a sentence which is neither a topic sentence nor a relevant sentence.

Crowd Agreement and Performance

To obtain the aggregated answers, we set 2 agreements as a threshold for identifying topic sentence and 3 agreements as a threshold for identifying (ir)relevant sentence, respectively. The detail performance of topic sentence, relevant sentence, and irrelevant sentence are described in Figure 3.4, 3.5, and 3.6, respectively.

T agree	R agree	precision	recall	f1 score
2	1	0.066667	0.031250	0.042553
2	2	0.150000	0.139535	0.144578
2	3	0.205882	0.325581	0.252252
2	4	0.158333	0.441860	0.233129
2	5	0.183007	0.651163	0.285714



Figure 3.6: Results of identifying irrelevant sentence by aggregating relevant keyword annotations with different threshold.

3.3.2 ML-Based Methods

To solve the problem lacking labeled data, we propose three naïve methods to predict topic sentence, relevant and irrelevance sentence.

Word Similarity

In our work, we choose two well-known methods for obtaining word similarity, Word2vec¹ and Wordnet². Word2vec is a group of related models that are used to produce word embeddings. It is trained by skip-grams and CBOW (continuous bag-of-words) and turns words into a vector; its pairwise similarity is got by cosine distance. Wordnet is a corpus built by experts; its pairwise similarity is got by path similarity. We use these two methods to define the distance of two words and the synonym between two words.

Topic Sentence Prediction

This section we purposed three kinds of method to predict topic sentence: rule-based method, TF-IDF (term frequency-inverse document frequency) and average sentence similarity. The notation s in below means a sentence, which is also a set of word w , and the set of sentences in a paragraph is noted as P . A represents the set of P , which means the whole article. The $length(s)$ represents the total word count in sentence s . The total relation between these sets is $w \in s \in P \in A$.

¹<http://deeplearning4j.org/word2vec>

²<https://wordnet.princeton.edu>

- Rule-based method: A paragraph usually begins with a topic sentence [37]. We adopted the first sentence rule in the rule-based method.

- TF-IDF: A topic sentence is a sentence that identifies the main idea in a paragraph. Each paragraph should have a different main idea in the standard essay writing. In other words, a topic sentence can be regarded as a sentence that contains the most number of keywords in the paragraph. Therefore, we adopted the concept of TF-IDF to extract keywords in the paragraph. Term frequency is calculated by sentence and inverse document frequency is calculated by paragraph. For term calculation, we aggregated the results of words with a high similarity. This method is used because with too few articles, and the same term may appear too few to find it second times. Sentence with highest average TF-IDF would be chosen to be the topic sentence in a paragraph.

$$tfidf_{w,P} = \frac{\max_{s \in P} \sum_{w \in s} tfidf_{w,s}}{\sum_{s \in P} \sum_{w' \in s} n_{w',s}} \times \log \frac{|P|}{|\{P \in A : w \in P\}|}$$

- Average sentence similarity (ASS): If there exists a topic sentence, it should have smallest average distance to other sentences in the paragraph. This method based on the corpus and calculate a pairwise distance of words between sentences in a paragraph, and then average the sum by the word aggregation. Sentence with smallest average distance sum would be chosen to be the topic sentence of a paragraph.

$$\min_{s \in P} \sum_{w_i \in s} \sum_{w_j \in P \setminus s} distance(w_i, w_j) / length(s)$$

Irrelevant Sentence Prediction

Based on the outcome of topic sentence prediction, we predicted irrelevant sentences by calculating the similarity of a sentence with a given topic sentence based on two kinds of corpus mentioned above. For Word2vec, we used “cosine similarity”; for Wordnet, we used “path similarity.” Then, we used leave-one-out validation to train PLA (perceptron

learning algorithm) for (ir)relevant sentence identification. The ground truths are annotated by two experts (as described in the following paragraphs).



3.3.3 Evaluation

To evaluate these methods, we recruited 15 participants who are all non-native speakers to write an essay in 30 minutes and crowdsourced those essays to obtain annotations from crowd workers. We compared precision, recall, and F1-score for evaluating the performance of prediction of topic sentences and irrelevant sentence.

The precision is the number of correct annotations divided by the number of all collected annotations. The recall is the percent of all correct annotations that are collected from an essay. The F-score that combines precision and recall is also used to evaluate the effectiveness of retrieved sentence.

Ground-Truth Data

We recruited two experts with 5+ ESL teaching and training experience to construct the gold standard annotations for 15 essays. Both experts have Ph.D. degree, and one's major is Applied Linguistics, and the other's is English Education. They were asked to annotate topic sentence, relevant keywords, and irrelevant sentence independently. The topic annotation (Kappa $k = 0.98$, $p < .0001$) and relevant keyword annotations (Kappa $k = 0.92$, $p < .0001$) created by two experts had high consistency. The gold standard annotations are the union of two experts' results.

According to our observations, the two experts followed consistent principles to annotate relevant keywords.

- identify supporting data with high relevance to the topic sentence
- identify the synonym appearing in the sentence with an argument but ignore simply repeating keywords
- identify chunks with a specific relation like cause-and-effect, etc.

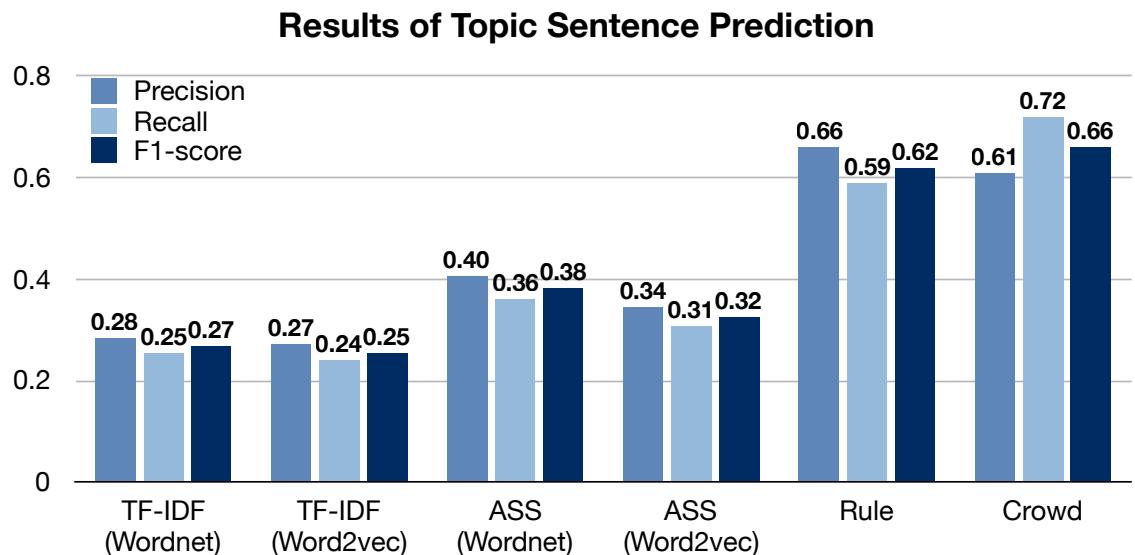


Figure 3.7: Results of topic sentence identification from ML-methods and our crowd-based method

Crowd Annotations

We crowdsourced 15 essays using our workflow in Amazon Mechanical Turk. 106 distinct workers were recruited for identifying topic sentence and relevant keywords. 55 workers completed 75 topic tasks (15 HITs with 5 assignments) and 51 workers completed 445 relevance tasks (89 HITs with 5 assignments). In total, there were 336 topic annotations and 1923 relevance annotations created in the workflow. The total cost is \$26.

Topic Sentence Prediction

Figure 3.7 shows the results of topic sentence prediction for three ML-based methods and the crowd-based method. The best performance is the crowd-based method (agreement=2). Its precision, recall and F1-score 0.61, 0.72, and 0.66, respectively. The worst result comes from TF-IDF with Wordnet, its precision, recall, and F1-score is 0.28, 0.25, and 0.27, respectively. The rule-based method (all-first) is slightly worse than the crowd-based method. In detail, it has higher precision but lower recall than the crowd-based method.

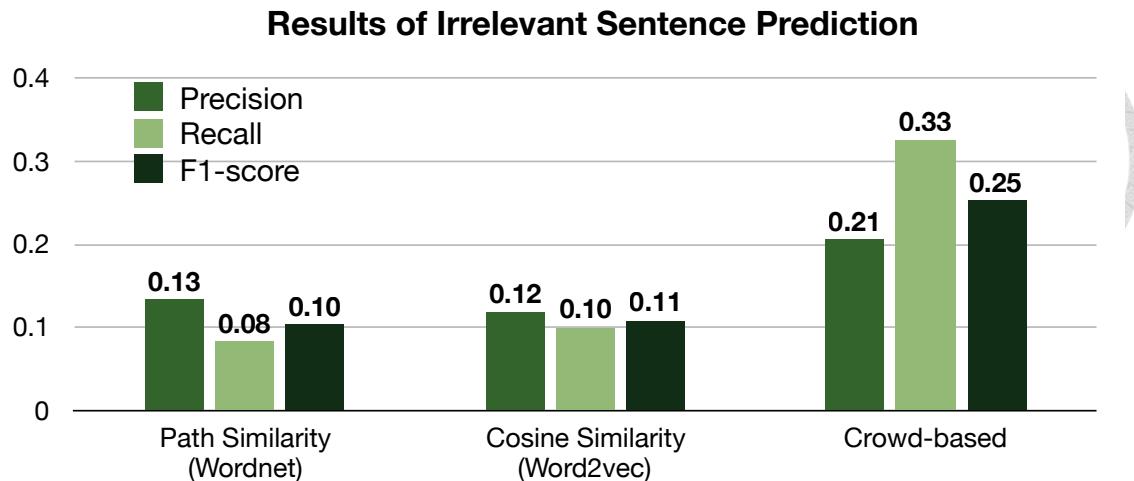


Figure 3.8: Results of irrelevant sentence identification from ML-methods and our crowd-based method

Irrelevant Sentence Prediction

Figure 3.8 shows the results of irrelevant sentence prediction. The crowd-based method (agreement=3) outperforms other methods on irrelevance sentence prediction. Its precision, recall, and F1-score is 0.21, 0.33, and 0.25, respectively. While the result of the worst one, the similarity with Wordnet, is 0.13, 0.08, and 0.10, respectively. The recall of Crowd is three times higher than similarity-based methods.

3.4 Field Deployment Study

To evaluate our system, we conducted a field deployment study for understanding how and whether StructFeed benefits ESL learners and helps them improve the quality of writing. In this study, 18 participants aged 19-35 years (56% male) were recruited from online postings on social media sites such as Facebook and Bulletin Board System. Each participant was self-motivated and had a common goal of practicing his/her writing skill. No compensation was provided to participants.

3.4.1 Study Design

We used a between-subjects study design and separated 18 participants into three groups. Three feedback-generation mechanisms were compared: expert feedback, free-form crowd

feedback, and StructFeed. The first mechanism is traditional writing feedback generated by one expert. Experts were recruited from Wordvice³, a professional online editing, and proofreading service. We used a particular service named TOEFL writing editing. The expert will edit and provide diverse feedback on the structure, content, grammar, and word choices of the article based on the grading rubric of TOEFL iBT. The second mechanism is free-form writing feedback generated by a single crowd worker recruited from Amazon Mechanical Turk. The crowd worker is asked to provide general writing suggestions about how to improve the unity and coherence of writing. The third mechanism is structural feedback with topic annotations, relevant keywords, and hints for writing generated by StructFeed.

3.4.2 Tasks and Procedure

The study consist of two tasks: writing task and rewriting task. Each task was performed on a different day. First, each participant was asked to perform a writing task in 30 minutes with one of three selected topic. The three selected topics were obtained from the list of past TOEFL independent writing questions, published by Educational Testing Service (ETS). Each group of participants generated an equal number of writing with the three different topics. Next, each participant was asked to revise their writing based on feedback obtained from one of the three feedback-generation mechanisms. After each task, participants were allowed to check grammatical errors of their writing with Grammarly⁴, a popular online grammar checker.

3.4.3 Measure

To measure the quality of the original and revised articles, we recruited two experts to rate all essays based on the writing scoring rubrics of TOEFL iBT. The rating scale is from 0 to 5 with a 0.5 interval. The higher the rating, the better is the quality of writing. Two raters independently rated all essays blind to condition. Ratings from the two raters were highly

³<http://wordvice.com>

⁴<https://www.grammarly.com>

	Feedback Generation Process			Revision Quality		
	Time	Quantity (number of suggestions)	Cost (USD)	Mean of Diff-Rating	# of Equal Diff-rating	# of Decreased Diff-rating
Expert Feedback	24~48 hrs	55.44	\$16.00	0.29 (SD=0.43)	1	1
Crowd Feedback	10~30 mins	8.11	\$2.00	0.38 (SD=0.44)	1	1
StructFeed	1~5 hrs	-	\$1~\$1.7	0.54 (SD=0.25)	0	0

Figure 3.9: The performance of three feedback-generation process is shown at the left three columns. The result of revision quality by three feedback-generation mechanisms is shown at the right three columns.

correlated, showing sufficient inter-rater reliability (Pearson $r = 0.93, p < .0001$). In the end, we averaged the ratings of the two raters to get a quality measure for each writing. We also calculated the difference between the average rating of the original article and the revised article to get a measure of rewriting performance for each participant.

3.4.4 Results

In this study, we first summarized the performance of each feedback-generation mechanisms by time, quantity, and costs. Next, we compared the mean of the difference of rating between original article and revised one, the number of equal rating, and the number of decreased rating. The results are shown in Figure 3.9.

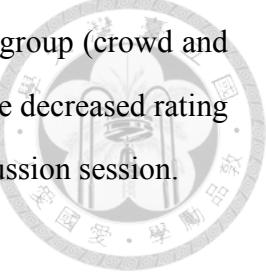
Time, Quantity, Costs

We usually spent \$16 to get expert feedback back in 24-48 hours, and got 55.44 comments (including grammar fix and suggestions about organization) for each article; we spent \$2 to get crowd feedback back in 10-30 minutes, and got 8.11 comments; we spent \$1-1.7 to get feedback by StructFeed in 1-5 hours.

Difference of Rating

We compared the mean of the difference of rating between original articles and revised articles for each group. We also investigate whether all participants improve the quality of writing after receiving feedback. People who received StructFeed got the best performance than other mechanisms. Every participant increased the quality of writing after

receiving feedback generated by StructFeed. Surprisingly, participants who received expert feedback got the worse performance. In addition, each free-form group (crowd and expert feedback) had two participants who had one equal rating and one decreased rating after revision. We will discuss this interesting phenomenon in the discussion session.



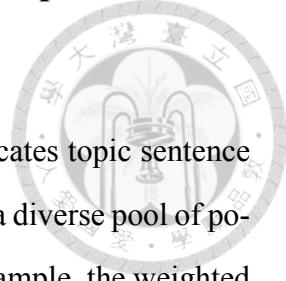
3.5 Discussion

3.5.1 Crowd helps develop better rules for machine

The results showed that crowd-based method outperformed all ML-based methods regarding identifying topic sentence and irrelevant sentence. According to our observation of aggregated crowd annotations, we found that the topic sentence of the introduction paragraph is usually the last sentence instead of the first one. Therefore, we modified the rule-based method to satisfy the new rule which the topic sentence is the last sentence in the first paragraph and the first sentence in the other paragraphs. The new rule derived from crowd annotations outperformed all initial methods. The precision, recall, and F1-score is 0.81, 0.72, and 0.76, respectively.

The obtained results also corresponded to the findings drawn from the interviews. Many participants reported that they would follow “four-paragraph essay template” to write a TOEFL independent essay in 30 minutes; besides, few participants who lacking knowledge about essay writing still kept putting their topic sentence at the beginning of all paragraphs. Therefore, the results suggest that crowd-based method is the better choice to analyze ESL writing which may be poor-structured and high diversity. Furthermore, the ultimate goal of proposed framework is to enable sustained crowd-machine collaboration. The writing support system and the ESL writer can gradually improve their skills together.

3.5.2 StructFeed not only identifies writing issues but promotes reflection



Unlike the automated methods, StructFeed not only identifies and locates topic sentence and irrelevant sentence but also provides diverse perspectives of how a diverse pool of potential readers (i.e., crowd workers) may interpret the writing. For example, the weighted annotations help people understand the gap between their intentions and readers' interpretations. Participant P7 reported that StructFeed helped her realize that the example she used in her essay might confuse other people and she said she would choose a more suitable example in her further revision.

Furthermore, it is clear that StructFeed is more flexible than ML-based method because it can be applied to support the writing of different topics and genres without the needs of training, the availability of corpus or prior knowledge for composing decision rules.

3.5.3 Expert feedback performed worse than crowd feedback?

It is indeed a surprising result. While we don't have hard proofs yet, here are some conjectures from our observations.

- Each revision is limited to 30 minutes. Under the pressure of time and overwhelmed by a large number of editing/comments from the expert, the ESL writers could take the easy way out by simply clicking to accept the suggested editing without making revisions on their own.
- Expert feedback contains both global issues such as lack of structure or coherence and local issues such as grammatical fix, word choice, etc. The ESL writers tended to focus on the easier fixes of local issues, rather than making the more difficult ones (global issues).
- There might exist the knowledge or communication gap between an expert reviewer and a novice writer. Some participants reported that they want to communicate with experts for further clarifications.

To solve this issue, we are now exploring how to present or filter the comments for facilitating better revision in our ongoing project.



3.6 Conclusion

In this work, we present StructFeed, a system that helps ESL writers to improve the quality of writings by receiving structural feedback. In the system, a crowdsourcing workflow was proposed to guide crowd workers to annotate topic sentences and relevant keywords through micro-tasks. By aggregating crowd annotations, the system can generate writing hints for directing people to address the structure issues effectively. In a field deployment study, we showed that our system could help ESL writers improve their writings. In addition, people who received feedback from StructFeed outperformed people who received feedback from a expert or a crowd worker. StructFeed enables new kind of writing feedback that cannot obtain from other sources. The work pioneers the design space of generating writing feedback with crowdsourcing mechanisms for ESL writers.



Chapter 4

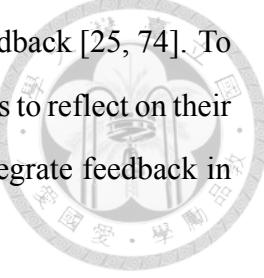
Feedback Orchestration: Supporting Reflection and Awareness in Revision

4.1 Introduction

Writing is rewriting. Revision has been established as one of the most important and complicated components in the writing process [21, 23]. Effective revisions rely on high-quality feedback, and the writer is expected to develop a revision plan by addressing those issues in some order [23]. Multiple revisions are often necessary to fix writing errors locally and to improve content or structure globally. Triggered by different types of feedback, each revision consists of a series of micro-tasks, such as adding examples to support a claim or fixing grammatical errors. However, novice writers tend to focus mainly on surface revisions because they struggle with developing good strategies to deal with structural problems [66, 18, 23].

This work addresses the challenge of supporting novice writers to take advantage of feedback and solve writing issues in their revisions. Previous research has used Online Feedback Exchange systems [24] to collect effective feedback to help designers revise their creative work. While most research focused on improving the diversity and quality of feedback [72, ?, 46, 27, 75], the difficulty in integrating feedback into revisions efficiently and effectively has been largely neglected. Recent studies have noticed the critical

gap between feedback provided and the subsequent performance [?, 25, 74] and start to develop approaches that help novices enhance the understanding of feedback [25, 74]. To fill the gap, our research explore a new way that facilitates novice writers to reflect on their writing with feedback and guides them to select good strategies to integrate feedback in revisions.



A formative study was conducted to understand how novice writers integrate feedback into revisions. The first finding was that inexperienced people usually revised in an unstructured way. For example, they used the most intuitive but inefficient approach to edit from beginning to end. They tended to focus on solving easier or specific local issues such as grammar and mechanics instead of more difficult ones such as argumentation and organization. The second finding was that writers having varying background and experience developed distinct revision strategies. Our study and research both suggest that novices need more supports to enhance their abilities and awareness as a writer to achieve structural revision, which is a pattern that experienced writers usually perform [31].

To support effective revisions, we propose *Feedback Orchestration*, which uses a rhetorical structure to guide novice writers interpret and implement feedback in a flexible revision workflow (see Figure 4.1). It enables writers to think and revise structurally based on structured feedback. Moreover, we present ReviseO, a system that implements the concept of *Feedback Orchestration*. It provides a standard writing rubric to help writers assess the writing goals with multiple criteria. Each feedback is classified into three categories by an automatic method. The system also enables three types of revision workflows (high-to-low, low-to-high, and all) for supporting writers to solve writing issues by utilizing categorized feedback.

In this work, we ran a field study that allowed writers to revise self-written articles with our system. We used a within-subjects study design to evaluate the perceived usefulness of the system, and also explore how three standard feedback presentation strategies (high-to-low, low-to-high, and all) affect revision behaviors and task performance. In the study, twelve self-motivated English as a Second Language (ESL) writers were recruited to revise three independent essays of 300-400 words based on structured expert feedback

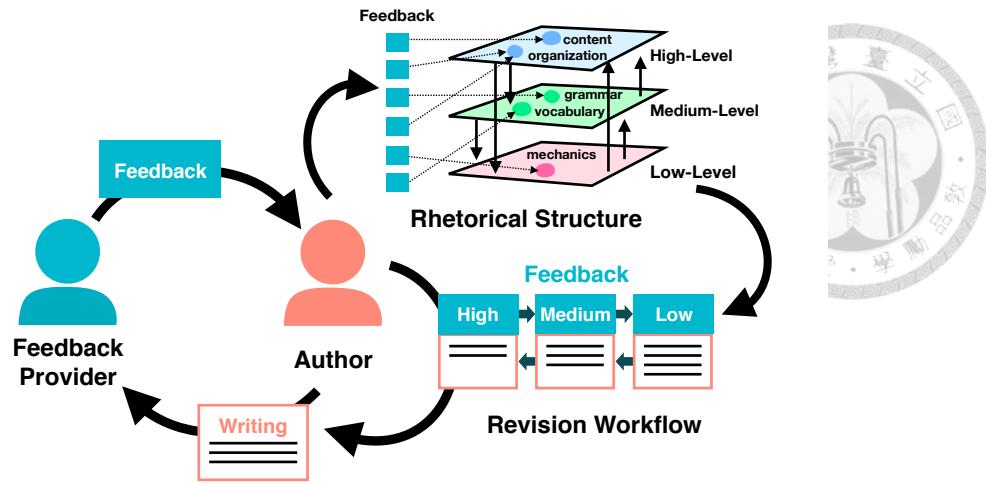


Figure 4.1: The overview of feedback orchestration. It guides writers to integrate feedback into revisions by a categorized structure.

using three different feedback presentation strategies. We present the findings based on analysis of data with interview transcripts and suggest design implications of online feedback system for writing supports.

The contributions of this work are:

- The first formative study contributing empirical understandings of writers' revision behaviors while dealing with unstructured feedback in revisions.
- The concept of Feedback Orchestration that uses a rhetorical structure to guide effective revision process by orchestrating feedback of different type.
- The system ReviseO, which supports three standard revision workflows for helping writers resolve issues by receiving expert feedback structured by a standard rubric.
- Results showed that structured feedback helped writers identify weaknesses and promoted reflection. Furthermore, novice writers who experienced three types of workflow increased their awareness and developed good strategies.

The insights obtained from this work help further researchers of related fields understand writers' revision behaviors while using different feedback strategies and contribute to the design of a flexible feedback framework for writing support.

4.2 Formative Study

To understand the challenges people commonly face while implementing feedback into revisions, we conducted a pilot study that allowed novice writers to revise an essay based on received feedback. The goal of this study is to observe how they revise and what kinds of strategies they adopt to deal with writing problems by utilizing generic feedback.

Six participants containing one female and five males are recruited. They aged between 18 and 23 years old; four are college students and two are graduated students. All of them are English as a Second Language (ESL) students with intermediate or upper-intermediate English proficiency and less than 1 year English writing experience. The reason we chose non-native English writers is that most of them are novice writers and struggle with improving writing based on feedback. As extreme users, their needs are amplified and their workarounds are often obvious. Revision work for non-native English writers is more complicated because they need to deal with not only the revision challenge but also the linguistic challenges of writing in a second language [31].

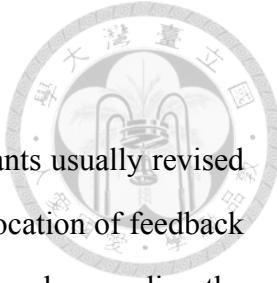
4.2.1 Task and procedure

The study consists of three stages: 1) read an essay, 2) revise the given essay with received feedback, and 3) take a post-task interview. First of all, participants read an essay written by other people. This essay consisted of multiple types of writing problems including high-level (e.g., content and structural issues) and low-level issues (e.g., grammatical and vocabulary issues). Second, participants revised the essay based on feedback in 30 minutes. Next, they took semi-structured interviews and answers several questions regarding how they dealt with the comments and what kinds of revision strategy they applied to their revision.

To examine how novice writers revise while receiving mixed types of feedback, all participants were asked to revised the same essay based on the same feedback generated by an expert, who was recruited from Wordvice¹, a professional editing service. The form of feedback consists of direct editing suggestions and written comments.

¹<http://wordvice.com.tw/toefl-writing-editing/>

4.2.2 Findings



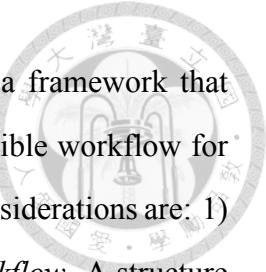
The results suggest four main findings. First, inexperienced participants usually revised in an unstructured way. They dealt with writing issues based on the location of feedback in the article. For example, they fixed each writing issue line by line when reading the article with feedback from beginning to end.

Second, inexperienced participants tended to focus on low-level feedback instead of high-level one. Even though they had been aware of the critical impact of high-level feedback on writing quality, they usually ignore it and failed to implement it into revision. Some participants were not accustomed to the concept of multiple drafts [47], and they naturally viewed revision as a means to “correct surface errors”, without trying to develop and refine content.

Third, most participants tended to give priority to dealing with comments with high-specificity and low-complexity. A comment with high-specificity indicates a comment not only describes writing problems but also suggests possible solutions or specific editing steps. For example, people would quickly accept or reject some direct editing suggestions made by reviewers, for it is easier to take action comparing with some judgments like “unclear sentence” or “awkward word.”

In the end, we found that people adopted different revision strategies while they received mixed types of feedback. Some people preferred to classify comments and then used high-to-low or low-to-high strategies to deal with feedback of different levels. This observation was aligned with writing literature, which has also suggested writers to fix high-level issues before low-level issues in general situations [60, 57], but adopt a reverse strategy when low-level issues damaging the global meaning of idea [3]. Therefore, we need to develop a suitable way to guide novice writers to think and revise structurally and encourage them to deal with high-level issues. A flexible revision process is also needed to support varying revision strategies.

4.3 Feedback Orchestration



To enable effective revision, we introduce *Feedback Orchestration*, a framework that guides writers to reflect and revise structurally and also support a flexible workflow for writers to apply their preferred revision strategies. Three key design considerations are: 1) *a rhetorical structure*, 2) *meta-feedback*, and 3) *a flexible revision workflow*. A structure is used to support writers interpret feedback and assess their work based on the standard criteria. Meta-feedback is a rhetorical label that helps writers be aware of the relationship between feedback and standard criteria. A flexible revision workflow enables writers to think and revise structurally by applying their preferred strategies.

Figure 4.1 show the overview of Feedback Orchestration. First, a writer submits the writing and requests for feedback from feedback providers. Next, feedback providers generate feedback based on the standard criteria. Each feedback is classified into rhetorical categories. Most important of all, a writer is guided to revise the work stage by stage through a revision workflow.

4.3.1 Expert revision practice

This framework draws inspirations from expert practice and the theory of revision from writing literature. Experts start with defining the rhetorical problems, set the goals and break them into sub-goals; then, they continuously reflect and revise the work based on the writing goals [22]. The revision process consists of three main stages: 1) detecting problems in a text, 2) diagnosing those problems, and 3) selecting a strategy to deal with those problems [23]. Experienced writers tend to move freely through these different stages and revise in a recursive, hierarchical manner; however, inexperienced writers struggle with detecting and diagnosing problem, as well as developing a good strategy for dealing with high-level problems [66, 23].

Research suggest that expert writers have three kinds of awareness (metarhetorical, metastrategic and metalinguistic) and four kinds of skill (collaboration, genre, text and context, and tools), leading them to revise in the very sophisticated way [31]. In particular, Horning emphasizes that these three kinds of awareness is the key for expert writers

to reflect on high-level writing goals and revise the work by breaking it into sub-tasks. Metarhetorical awareness is the awareness or knowledge of one's self as a writer, including typical strategies and approaches to writing and revising. It has a substantial impact of their ability to revise effectively. Metastrategic is the awareness or knowledge of their own personality type and its influence on their writing or revision behaviors. Metalinguistic is the awareness or knowledge of terminology to discuss language issues. Therefore, this framework has to support reflection and awareness, leading novice writers to revise effectively.

4.3.2 Support reflection and awareness

Research suggest that managing writing with an eye for rhetorical categories (mechanics, organization, semantics) has been shown to support the writing process, especially for novice writers [58]. Therefore, we decide to apply a rhetorical structure to scaffold novice writers to assess their writing with the standard criteria and meet the rhetorical goals, which is the activity that expert writers always perform in the mind. In addition, a rhetorical label as meta-feedback is assigned to each feedback. It is used to help novice writers be aware of the relationship between feedback and the rhetorical goals. In the end, structured feedback can be used in a flexible workflow for helping writers think and revise in a structured way.

4.4 System Implementation

This paper design and implement ReviseO, a writing support system that demonstrates the concept of Feedback Orchestration. In the current implementation, we use experts as feedback providers to generate high-quality feedback in order to reduce the influence of the quality of feedback on revision outcome. We will introduce which structure is selected, how feedback is classified into the rhetorical categories, and what kinds of revision workflow we attempt to explore, and how we enable a flexible revision workflow by structured feedback.

4.4.1 Taxonomy of writing feedback

Learning science has suggested that using rubrics can help students self-assess their own work and improve learning and performance [35, 59]. Well-designed rubrics can help students align their own work with concrete criteria, guiding them to better grasp a domain-specific knowledge and key principles [6]. Therefore, we plan to use well-designed rubrics as a structure to classify feedback.

To understand what kind of feedback is important for improving writing quality, we consulted several English as a Second Language (ESL) writing instruction books and experienced ESL teachers in writing center. We follow Jacobs *et al.* [34] ESL Composition Profile, which is the best-known evaluation criteria used in the ESL writing domain, to classify feedback into five categories: content, organization, vocabulary, language use, and mechanics.

Content refers to the substance of writing, for example, the main idea or supporting details; organization refers to the logical organization of the content; vocabulary refers to the selection or words those are suitable with the content; language use refers to the use of the correct grammatical forms and syntactical pattern; mechanics refers to all the arbitrary “technical” stuff in writing like spelling, capitalization, punctuation, and use of numerals and other symbols. The detailed definition and example of each category in the taxonomy are depicted in Figure 4.2.

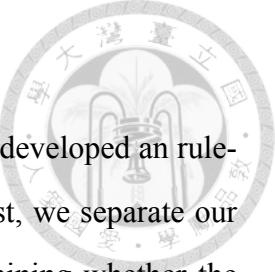
In this work, we grouped five categories into the three broader categories of rhetorical feedback (content and organization) as high-level feedback, linguistic feedback (vocabulary and language use) as medium-level feedback, and mechanics as low-level feedback. This grouping follows a standard classification in writing center literature [60], which group content and organization together as “higher-order concerns” and grammar and vocabulary together as “lower-order concerns”. The only difference is that we separate the mechanics as a category from “lower-order concerns,” for most of this type of errors can be corrected by automated proofreading tools. In our system, an automated proofread checker can be further included for solving this type of errors.



	Feedback type	Definition	Examples
High	Content	Content refers to the substance of writing. It includes topic sentence expressing main argument and supporting ideas.	e.g. unity of argument, supporting idea, relevant example, addresses the question, etc.
	Organization	Organization refers to the logical organization of the content.	e.g. coherence of the content, relation between sentences, logical sequencing, etc.
Medium	Vocabulary	Vocabulary refers to the selection of words those are suitable with the content.	e.g. word choice, etc.
	Language use	Language use refers to the use of the correct grammatical forms and syntactical pattern.	e.g. fixing grammatical errors, or paraphrasing, shortening, etc.
Low	Mechanics	Mechanic refers to all the arbitrary technical stuff in writing like spelling, capitalization, punctuation, etc	e.g. spelling errors, punctuation, capitalization, format, etc.

Figure 4.2: The taxonomy of writing feedback follows the ESL Composition Profile. It consists of five writing evaluation criteria including content, organization, vocabulary, language use and mechanics. We grouped content and organization as high-level feedback, vocabulary and language use as medium-level, and mechanics as low-level feedback.

4.4.2 Feedback classification



To deal with the large quantity of feedback received from experts, we developed an rule-based classifier to categorize each feedback into four categories. First, we separate our feedback into two categories, comment and modification, by determining whether the feedback contains direct editing suggestion. *Comment* is the type of feedback with suggestive or critical opinions. For this kind of feedback, we used natural language processing toolkits (NLTK) to analyze the sentiment of the comments and Rapid Automatic Keyword Extraction algorithm [62] to extract the keywords from the comments. *Modification* contains all types of direct edition suggestions like deleting, inserting and replacing words and characters. For this type of feedback, we pass them to the next step to classify them into three groups: mechanics, grammar and vocabulary. Mechanical errors consist of punctuation and typographical errors appeared in articles. And since our original articles had fixed their typographical errors with grammar checkers, the mechanical errors are mainly punctuation errors and capital and lowercase mistyping. In our system, we determine modification as mechanical errors by detecting whether the feedback contains only punctuation revising, capitalization or lowercase.

After classifying mechanical errors, we then separate the grammar errors from the feedback. Grammar errors are errors like misusing of singular/plural or tenses. The common point of grammar errors is that the modified text has the same stem with the original text. Therefore, we use stemming and conjugating provided by NLTK to determine whether the feedback is a grammar error. Vocabulary errors are comparatively complicated. In our system, if the modification is neither in the category of mechanical errors or grammar errors, then we determined it as a vocabulary errors.

We evaluated our method with 21 articles. Each article had an average of 83.43 suggestions. For three categories (high-level, medium-level, low-level), the average error rate was 0.017. For four categories (comment, grammar, vocabulary, mechanics), the average error rate was 0.034.

4.4.3 Revision workflow

The formative study suggests that writers adopted varying editing strategies on their revision based on different context and their writing expertise. According to writing literature. There is also still no conclusive evidence to indicate which editing strategy is the good one. Most of ESL writing literature suggest that people should consider global issues before local issues [60, 57]. On the other hand, Blau *et al.* suggest that addressing local errors before global ones can be useful and productive when the writing's clarity is compromised by many local errors [3]. However, the common suggestion made from the literature is that dividing writing issues into multiple groups and fixing them separately.

In this work, we attempt to enable a flexible revision workflow based on structured feedback. Writers can apply different revision strategies to focus on dealing with one type of writing issues and then move to solve another type of issues.

This system currently support three types of revision workflow: (1) showing categorized feedback together (ALL), (2) showing categorized feedback in high-to-low order (HML), and (3) showing categorized feedback in low-to-high order (LMH). Writers are allowed to process feedback together or in high-to-low and low-to-high order. The high-to-low revision workflow allows writers to receive high-level feedback like content and organization first, and the lower level of feedback like language use, vocabulary, and mechanical errors later; the low-to-high revision workflow allows people receive different types of feedback in a reverse way.

4.4.4 Revision interface

In the system, writers can obtain a number of categorized comments located in their original articles. Writers can execute three main operations including rating feedback, accepting or rejecting feedback, and revising the article in the editing area. The revision interface shows the original article view (a), and editing area (b) on the left side. The feedback area is located on the right side. Each comment has its own rating (c), and accept and reject buttons (d) (see Figure 4.3).

In the default setting (the concurrent editing mode), the interface shows all categorized

Time spent: 1 min 49 sec

What gift would you give to help a child develop?

The gifts I would give to help a children develop include a bike and a series of storybooks. Below I would explain the reasons about why the selection as is and how the the reasons connect to my personal experience.

The selection of bike is for having him/her gain **a** motor ability in bike riding as well as the capability for exploring a wider world. In my childhood, my older brother gave me a bike as a birthday gift when I was 8. To be able to ride it smoothly, I practiced a lot, and finally got the skill. Though I fell to the ground several times and got harmed during the process, I was happy and found it worthy. I rode my little bike to explore the neighborhood I lived and got a lot of surprise.

Editing Area for revision #1:

The gifts I would give to help a children develop include a bike and a series of storybooks. Below I would explain the reasons about why the selection as is and how the the reasons connect to my personal experience.

The selection of bike is for having him/her gain **b** motor ability in bike riding as well as the capability for exploring a wider world. In my childhood, my older brother gave me a bike as a birthday gift when I was 8. To be able to ride it smoothly, I practiced a lot, and finally got the skill. Though I fell to the ground several times and got harmed during the process, I was happy and found it worthy. I rode my little bike to explore the neighborhood I lived and got a lot of surprise.

Another selection, a series of storybooks, is for insure the child's imagination. From my personal

Comment.
Organization: Excellent. Your points are set out very well, followed a clear thesis, and ended on a strong note.

Accept Reject

Content & Organization #2



Comment.
Content: Very strong, the points made were detailed, concise, and salient.

Accept Reject

Content & Organization #3



Comment.
Very literal wording, which draws attention to the fact that this is an essay. More natural and academic wording is recommended.

Accept Reject

Next



Time spent: 2 min 20 sec

What gift would you give to help a child develop?

The gifts **I** would give to help a **children** develop **include** a bike and a series of storybooks. Below I **would** explain the reasons **about** why **the selection as is** and how **the the reasons** connect to my personal experience.

The selection of **bike** is **for having him/her gain the** motor ability **in** bike riding as well as the capability **for exploring** a wider world. **In my childhood**, my older brother gave me a bike as a birthday gift when I was 8. To be able to ride it smoothly, I practiced a lot, and finally got the skill. Though I fell to the ground several times and got harmed during the process, I was happy and found it **worthy**. I rode my little bike to explore the neighborhood I lived and **get** a lot of surprise.

Editing Area for revision #2:

The gifts I would give to help a children develop include a bike and a series of storybooks. Below I would explain the reasons about why the selection as is and how the the reasons connect to my personal experience.

The selection of bike is for having him/her gain the motor ability in bike riding as well as the capability for exploring a wider world. In my childhood, my older brother gave me a bike as a birthday gift when I was 8. To be able to ride it smoothly, I practiced a lot, and finally got the skill. Though I fell to the ground several times and got harmed during the process, I was happy and found it worthy. I rode my little bike to explore the neighborhood I lived and got a lot of surprise.

Another selection, a series of storybooks, is for insure the child's imagination. From my personal

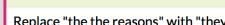
Language use #8



Replace "the selection as is" with "I would make this selection".

Accept Reject

Language use #9



Replace "the the reasons" with "they".

Accept Reject

Language use #10



Insert "a".

Accept Reject

Language use #11



Replace "for having" with "so that the child can".

Next

Figure 4.3: The two images shows the high-to-low interface that transits from the high-level stage (left) to the medium-level one (right). (a) original article view, (b) editing area, (c) feedback rating, (d) accept and reject buttons, (e) writing tip button, (f) next button, and (g) selected annotation.

comments together on the feedback area. The high-level feedback (Content and Organization) is displayed in light blue; the medium-level feedback (Vocabulary and Language use) is displayed in light green; the low-level feedback (Mechanics) is displayed in pink. By assigning a particular editing strategy, the interface only shows one type of feedback in each stage.

Figure 4.3 presents the two stages of a sequential revision workflow with a high-to-low editing strategy. The left interface only shows the content and organization feedback in light blue and switches to the right interface when pressing the next button (f). When writers click on the annotation on the original article, the annotation will change the color to deep pink, and the corresponding comment with a connecting line will show up in the interface (g).

In the whole revision process, writers can look up the important knowledge about writing by clicking the top-right button named “writing tip” (e) on the header of the interface. The writing tip explains the basic concept of essay structure, writing evaluation principles, and feedback categorization (see Figure 4.4).

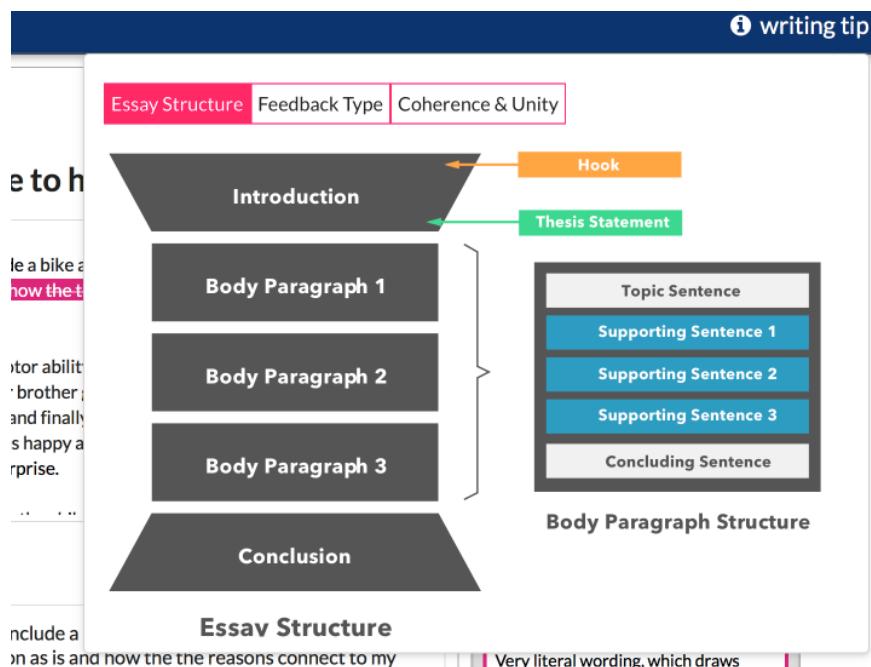


Figure 4.4: Writers can find helpful information about writing by clicking the top-right button “writing tip.”

4.5 Experiment

ReviseO hypothesizes that providing a feedback structure and three types of revision workflows can allow novice writers to be aware of the writing goal, reflect on their writing abilities and facilitate effective revision behaviors. In this experiment, we evaluated overall usefulness and helpfulness of the system, helpfulness of feedback categorization, and actual revision performance when learners used our system to revise their writings. Furthermore, we adopted a within-subject design for understanding writers' revision behaviors and experience when receiving categorized feedback in three types of revision workflows. Three experimental conditions are compared: (1) showing categorized feedback together (ALL), (2) showing categorized feedback in high-to-low order (HML), and (3) showing categorized feedback in low-to-high order (LMH).

In the experiment, we recruited self-motivated learners to edit their writings with the ReviseO system. Each participant was asked to write three essays and received feedback generated from an expert. When they performed their revision tasks, they experienced a different experimental condition for each essay. To reduce the learning effect, we counterbalanced the order of experimental conditions.

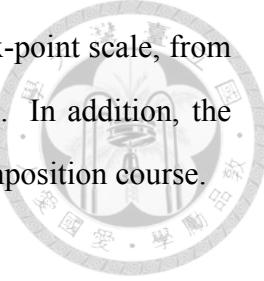
4.5.1 Participants

We recruited 12 volunteers (6 females and 6 males) aged 22 to 34 through online community postings in Facebook and a Bulletin Board System. All participants were college educated at the bachelor degree level; two of them were at the master degree. All participants were EFL learners located in Taiwan and were self-motivated to participate in the experiment with a goal of polishing his/her writing skills or preparing for the TOEFL exams.

Before the experiment, participants were required to take a English proficiency test and self-report their writing course experience. The English proficiency test we used is Cambridge English Proficiency Assessment² and the test scores were mapped to the level

²<http://www.cambridgeenglish.org/test-your-english/general-english/>

of the Common European Framework of Reference for Languages (CEFR)³, which is an International language standards. It describes language ability on a six-point scale, from A1 for beginners, up to C2 for those who have mastered a language. In addition, the writing experience is the number of year of taking English writing composition course.



4.5.2 Task and procedure

The experiment consists of two sessions: writing and revision. Participants who had finished the writing tasks were allowed to participate in the revision tasks.

Writing task

In the first session, participants received a request for completing three essays with assigned topics and submitting them by email. The three topics selected from a TOEFL writing collection included: (1) The best way of learning, (2) Contributions of artists vs. scientists, and (3) What gift would you give to help a child develop?

Participants were asked to write each essay in 30 minutes and correct basic grammatical errors in their writing with Grammarly, a popular online grammar checker, before submitting their writing. After submitting three essays, participants were invited to join the online experiment for revising their three essays with categorized expert feedback with the ReviseO system.

Expert feedback for the writing task

To obtain high-quality feedback, we use a particular service named TOEFL writing editing in Wordvice. This online service provides detailed and diverse feedback on the structure, content, grammar, and word choices of the article based on the TOEFL grading criteria.

Feedback collected from Wordvice included a score, overall feedback, direct editing and comments (see Figure 4.5). We used a Python script to extract all editing history and comments from a DOC format. In the preprocessing stage, the score is removed to prevent participants from ignoring high-level issues [68]. The overall feedback was separated into

³<http://www.cambridgeenglish.org/exams-and-tests/cefr/>

4 types of comments. Next, each edit was was considered as one suggestion and identified as insert, delete, replace, or comment. In the classification stage, each suggestion is classified by an automated classification method. In addition, all categorized suggestions would be converted to a JSON format and feed into the ReviseO system. Finally, the system can fully present all categorized suggestions with their belonging categories in different colors.

Contributions of artists vs. scientists

People often possess different perspectives on the issue of the contributions of artists and scientists. Some people **content** contend that artists contribute more, since they bring **creativity** to the society, **while** others think that scientists contribute more because they bring convenience to the world. **Base**-Based on my personal **experience** **experiences** and **observation** **observation** in life, I believe artists **have**-make more contributions to the world.

First of all, people live under **lots** of **great** stress nowadays, and artists can help **release** **reduce** **their** pressure. For example, artists **creative** **create** **fabulous** **music** which **that****make** **makes** **people** **feel** **relax** **relaxed** while listening. Whenever I have **lots** of **much** work to do, I listen to soft music **and** **indulge** **immerse** **myself** in another world, and feel **totally** **relax****relaxed**. Then I **refill** **the** **energy** **This** **reenergizes** **me** so that I **can** **and** **return** **go** **back** to work with higher efficiency. On the other hand, the high-technology **device****devices** invented by scientists **usually** **are** **used** **as** **a****tool****tools** **for** **work**. **Therefore** So, as long as we have them with us, we **feel** **get** **more** **stress**. Therefore, artists contribute more than scientists since their work can **release** **reduce** **people** **people's** **stress**.

Secondly, people put **an** **emphasis** on **money**, and **artist** **artists** **can** **bring** **extra** **revenue** **to** **a** **city****cities**. Some of the **incredible** **work** **works** made by famous artists are so attractive that people **from** all over the world want to **see** **it** **them** in person. When people **come** **to** **visit** **the** **a** **country** **for** **taking** **a** **look** **at** **see** the **great** **work****art**, they buy **food** **to** **eat**, **buy** **a** **beverage** **beverages**, **to** **drink** and **buy** **toys** **to** **entertain** **themselves****entertainment**. Hence, more money is brought into the city from other countries, which **increase****increases** the revenue of the city. As a result, artists can make money for their country by attracting tourists.

Last but not least, people care about **creativity**, and artists can help us broaden our **imagination****imaginations**. For instance, there are creative and pretty **beautiful** **statues** **made** **by** **artists** in **some** **famous** **parks** **made** **my** **artists**. People who visit the **park** **parks** can see **those** **these** **statues** and guess the story of the artwork. Some of the statues can even interact with people and surprise them. By doing **so** **Thus**, people get the chance to come up with **creative**, abstract ideas that they have never thought about with their **creativity**. **Therefore** So, artists **bring** **make** **more** **contributions** to the world by making people **picture** **interesting** **things** **with** **through** **special** **artwork** **artwork** **display** **displayed** **in** **public** **place****places**.

To conclude, although scientists can make our **life** **lives** **more** **easy****easier** and **more** **convenient** by inventing new equipment, I believe artists contribute more since people are **stressful** **stressed** and need a way to **release** **reduce** **their** **pressure**. **Also****Furthermore**, people care about **creativity** and **money** a **great** **deal**.

Score 2.5/5
Organization: Good. Your points are set out quite well.
Content: OK. You include some relevant arguments, but could consider making (and responding to) a counter argument. In addition, your conclusion is quite weak – it simply repeats points rather than adding to them.
Vocabulary: OK. Your language is generally relevant to the topic, but you sometimes use inexpressive or unnatural phrasing that makes your sentences awkward (see comments).
Grammar: OK. Your sentences are comprehensible, but often quite poorly or unnaturally phrased and/or syntactically incorrect (again, see detailed comments).
Please feel free to request me by name (Jayne) should you require an editor in the future. Thank you for using wordvice.com.tw.
Show less
From imported document

Replace: "content" with "contend"
From imported document

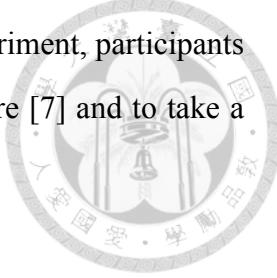
Figure 4.5: One example of expert feedback collected from one expert by using Wordvice TOEFL writing feedback service. The feedback include one score, overall feedback and specific feedback containing direct editing and comments.

Revision task

In the second session, each participant was scheduled in three different time slots for performing revision tasks using three types of revision workflows (all, high-to-low, low-to-high). Before the revision tasks, they were asked to take a English proficiency test and 5-minute writing training. The English proficiency test is used to measure participants' English proficiency level; the writing training is used to introduce the five concepts of the feedback categorization, the High-Medium-Low structure and how to use the interface for their revision.

In each revision task, participants were assigned one articles under a given conditions. They were informed that they would experience what kinds of feedback order and focus on

fixing one type of writing issues by stage. After finishing each revision task, participants were asked to answer a post-task questionnaire. After the whole experiment, participants were asked to answer the System Usability Scale (SUS) questionnaire [7] and to take a 15-minute interview.



4.5.3 Measures

Time spent and edit distance

We used time spent and edit distance between the original and revised essay to measure writers' explicit revision effort. We used the Levenshtein distance (LD) to calculate the similarity between two texts [42]. The value of distance is the minimum number of editing operations (insertion, deletion, and substitution) required to transform one string into another string, which is commonly used to measure changes between text in studies about writing feedback (e.g., [16], [54]).

Quality of improvement

To evaluate the quality of writing, we recruited two experts, one has more than five years of ESL teaching and the other has more than five years of professional editing experience, to rate all original and revised article. Two experts independently rated articles blind to conditions based on the ESL Composition Profile [34]. Ratings from the two experts were highly correlated, showing sufficient inter-rater reliability (Pearson $r = 0.93$, $p < .0001$). For each writing, ratings from the two raters were averaged to represent the quality for five aspects including content, organization, vocabulary, language use, and mechanics. Next, we calculated the difference of rating between original and revised articles by subtracting the score of the original article from the score of the revised one.

Post-experiment survey

We used two types of post survey to evaluate the usefulness and helpfulness of the Re-viseO system under three experimental conditions and the overall system usage experience. The first survey, which is answered after each revision task, is used to evaluate

Measures	Indicate on a scale of 1 (strongly disagree) to 7 (strongly agree)
Easy of use	- I think this system is easy to learn. - I think this system is easy to use.
Helpfulness (Overall system)	- This system helped me revise article - This system helped me understand the way of evaluating the writing. - I would like to use this system to practice writing.
Helpfulness (Feedback categorization)	- Classifying feedback into three types of writing issues helped me revise article. - Categorized feedback helped me identify the writing issues in the article.

Table 4.1: Survey items for perceived helpfulness of overall system and feedback categorization.

	Overall System		Categorization	Writing Quality	Revision Effort	
	Easy of use	Helpfulness	Helpfulness	Diff of Rating	Time spent (sec)	Edit distance
LMH	5.58 (.43)	6.22 (.21)	6.58 (.23)	6.79 (.88)	1731.58 (147.81)	353.33 (34.68)
HML	4.92 (.47)	6.22 (.25)	6.50 (.26)	7.21 (.53)	1727.75 (204.05)	370.00 (60.85)
ALL	5.25 (.33)	6.25 (.18)	5.67 (.48)	7.25 (.44)	1612.92 (124.57)	444.33 (72.35)

Table 4.2: The means (and standard errors) of perceived usefulness and helpfulness, as well as objective measures in each condition. There is no significant relationship between three conditions.

perceived usefulness and helpfulness by a list of questions on a 7-point Likert scale (see Table 4.1). In addition, we asked the additional questions about helpfulness of feedback structure under three conditions. The second one is System Usability Scale (SUS)[7], which is only answered at the end of the experiment (i.e., after the third revision task) and calculate the final score of perceived usability.

4.5.4 Results

We reported the findings by combining analysis of data and interview results. A two-way mixed ANOVA is used to help analyze the result over three conditions. Table 4.2 shows the perceived usefulness and helpfulness, the quality improvement between the original and the revised articles, as well as the average time spent and the number of edits between each stage in each condition (ALL, HML, LMH).

Participants spend less time and edited more in the concurrent workflow

Participants spend less time and edited more in the concurrent workflow (ALL) than in the sequential workflows (HML or LMH) by average. Although some participants (P9, P13) reported the consistent finding about less-effort and easy-achievement in the concurrent

workflow, the results was no significant difference of the average time spent ($F(2, 22) = .28, n.s.$) and the number of editing ($F(2, 22) = .84, n.s.$) among three conditions.

All participants improved the quality of writing

All participants improved their writings in three conditions. Some interesting insights obtained by combining those data with interview results are described as in the following paragraphs, while there is no effect of different editing workflows (ALL, HML, LMH) on total quality improvement.

High perceived helpfulness and usefulness

According to the result of the post-questionnaire (Table 4.2), participants felt satisfied with the ReviseO system and the feedback structure provided. All ratings were higher than the average score (3.5). In addition, we obtained a good score above the average score in a SUS usability test ($M = 72.08, SE = 4.82$). It is interesting that participants can clearly distinguish the difference among three types of workflows, identify each pros and cons, and present their preferences, while there was no significant difference of perceived helpfulness and usefulness among three experimental conditions. We will discuss the reason in the discussion section.

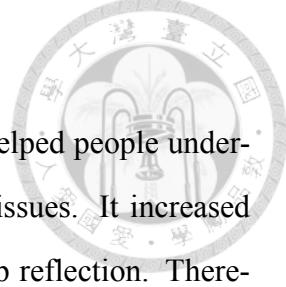
4.5.5 Insights from interviews with participants

Structured feedback promoted reflection

All participants reported that feedback categorization provided by the ReviseO system was very helpful for their revisions. We observe that feedback categorization can benefit people in three aspects, from filtering information to identify own strength and weakness to promoting learning behaviors. First, the feedback categorization helped filter information and discover specific writing issues.

“The categorization helps me quickly filter the information and I can focus on a specific type of writing issues instead of all of them. (P1)”

“It helps me understand the weakness of my writing and I realize which aspect of writing I need to improve. (P11)”



In addition, the feedback categorization offered a structure that helped people understand the relationship among the writing goal, criteria and writing issues. It increased people's awareness to identify common mistakes and supported deep reflection. Therefore, they can easily identify strengths and weaknesses of writing and develop a future plan for learning. Interestingly, one participant reported that understanding the current status of performance helped reduce their level of anxiety for learning. More research is needed to understand whether providing feedback categorization can reduce the learning anxiety.

“This categorized feedback helps me identify my common mistakes easily! When I see the same type of writing issues appearing frequently, I understand that I need to pay more attention to this type of problem in my next writing. (P2)”

“It helps me identify my weak points of writing. For example, the reviewer indicated that the structure of writing is good and there is no error in this category. I quickly move on to solve language problems without worrying about whether my writing exists unnoticed content and organization issues. (P11)”

Furthermore, the categorization provided new knowledge about how to evaluate writing quality and promotes active learning behaviors. People are motivated to apply the same structure to examine their own or others' writing in the future.

“I have never used such structure to assess my writings before. In this experiment, I learn new knowledge about writing from categorized feedback, and learn how to evaluate the quality of the writing by those good or bad examples in different criteria. I will use the same way to review others' writing next time!! (P7)”

“It helps memorize the key principles of the writing evaluation, and I will adopt it for examining my writing. Also, It may help me investigate others’ writing and learn from their strengths in different perspectives. (P13)”



Varying preference for revision workflows

People had varying preference to choose their revision workflow. Three participants (P2,P5,P9) preferred high-to-low workflow; four participants (P4,P8,P10,P14) preferred low-to-high workflow; two participants (P11,P13) preferred to receive all categorized feedback together. Interestingly, three participants (P2,P3,P14) developed new revision strategies after experiencing three default revision workflows and plan to use a “high-low-medium (HLM) strategy,” for their future revisions. These findings suggest that an effective feedback system should support varying revision strategies, which are aligned with our design considerations about flexible interface design. Our system design also support high-low-medium workflow, while it is not included in our experiment.

Furthermore, we found that non-novice writers (P2,P8,P11) tended to choose the workflow that allows them to apply consistent strategies and novice writers (P3,P4,P5,P9,P10,P14) tended to choose the different workflow, compared with their previous strategies (see Table 4.3). The results suggest that our system provides good opportunities for novice writers to explore different types of strategies using three types of revision workflows and helps them identify suitable revision strategies that they have never tried before. It helps improve meta-strategic awareness that novice writers usually lack.

Based on interview data, we summarized the reason why people had vary preference for sequential (HML or LMH) and concurrent workflow (ALL). People preferring the sequential workflow indicated that presenting feedback separately helps them focus on similar writing issues and resolve them together, and they complained that mixing all types of feedback led distraction and cognitive overload when switching back and forth between the high-level and low-level issue in the revision task.

“I like the design of 3-stage revision workflow. It allows me to focus on one type of feedback. In contrast, the interface showing feedback together is messy

ID	Score	English Level	Writing Experience (years)	Previous Strategy	Workflow Preference	Consistent	Note
P1	19	B2 (Low)	2 (Med)	Only-local	No pref	-	always ignore high-level issues
P2	17	B1 (Low)	3 (High)	Global-first	HML ->HLM	Yes	slightly modify previous strategy
P3*	18	B2 (Low)	0 (Low)	No exp	HLM	No	
P4*	16	B1 (Low)	1 (Low)	Global-first	LMH	No	
P5*	19	B2 (Low)	0 (Low)	Line-by-line	HML	No	
P7	21	C1 (Med)	0 (Low)	Global-first	No pref	-	no high-level issues
P8	24	C2 (High)	3 (High)	Local-first	LMH	Yes	
P9*	17	B1 (Low)	0 (Low)	Line-by-line	HML	No	
P10*	18	B2 (Low)	1 (Low)	Specific-first	LMH	No	
P11	23	C2 (High)	2 (Med)	Global-first	ALL	Yes	- reject misclassified feedback - want more control
P13*	18	B2 (Low)	0 (Low)	Line-by-line	ALL	Yes	- prefer less time and less effort
P14*	16	B1 (Low)	1 (Low)	Line-by-line	HLM	No	

Table 4.3: Non-novice writers preferred a workflow that is consisted to their previous revision strategies; novice writers changed their revision strategies after experiencing three types of workflows. The star symbol (*) indicates a novice writer who got low-level score and had less than one year of writing course experience.

so that I felt tired and impatient to see all of them. I need to scroll down the page for a long time to see all the comments. (P1)"

"They are messy. I felt exhausted after switching back and forth between high-level feedback and low-level feedback. (P10)"

"It is hard for me to find the information that I need to see when I obtain all types of feedback. It also makes me hard to recall what the writing problems I made in that experience. (P12)"

By contrast, people preferring the concurrent workflow (P11, P13) indicated that presenting feedback together is the most efficient way to revise the articles because they can follow their original ways to make revisions. That is to say, the concurrent workflow allows writers to freely revise article in any ways. It also decreases the influence of feedback categorization and novice writers would maintain the original revision behaviors and ignore high-level issues easily. In this experiment, one novices writer (P11) reported that he only cared language-related issue so he preferred a quick way for his revision. On the other way, The concurrent workflow support experienced writers to apply their strategies

in more flexible ways. For example, P13 mentioned that she mis-rejected the feedback in a high-to-low workflow, which is related to her previous revision strategy, so she thought that the concurrent workflow is much useful for her.



Sequential workflow guides writers to think and revise structurally

We found that sequential workflow can guide people to think and revise writing issues in a structure way. Participants who preferred high-to-low workflow indicated that using high-to-low helped focus on big picture first and examine the details later (P2,P5,P9). Particularly, P2 mentioned that the concept of high-to-low editing is related to their knowledge obtained in the previous writing course, and it is also consisted to his revision habit. In addition, people had more energy to deal with the harder problems at the beginning of editing process (P13).

By contrast, other people preferred low-to-high editing workflow, for it led to a smooth transition from one issue to another. They can start with small, easy problems and move on harder problems later. This process guided them to be familiar with the content and develop a deeper understanding of writing by fixing several small errors. These micro-accomplishments help people build momentum to solve harder problems on the same content.

“Either low-to-high or high-to-low workflow helps me develop a thinking order. It’s convenient for me to edit my writing. (P10)”

“I start with fixing language-level issues over the whole article and re-read my writing sentence by sentence. Then, I am getting more and more enjoyable because I become familiar with the content and easily recall the situation in which I was writing and understand why I generated those words. It helps me solve the structure issues more easily. (P4)”

Editing conflicts in the sequential workflow

Sequential editing workflow enhances the benefits of feedback categorization and facilitates structural revision behaviors. However, presenting feedback inappropriately may

lead to editing conflicts both physically and mentally. First, separating related comments into different stages may result in misunderstanding or confusion. The comments located in the same sentence may provide suggestions from different perspectives and they often required to be considered together. For example, one participant (P11) mis-rejected a suggestion about punctuation errors at the early stage because of an incomplete message delivered from only one comment. In addition, some participants (P1, P3, P11, P4) felt confused about comments suggesting to fix non-existent problems, for they had been solved at the previous stage. This conflict often happened in the high-to-low workflow because the scope of the high-level issue is larger than of the low-level issue.

“In the second stage, I rejected a comment suggesting an incorrect edit. However, I realized that the suggestion is correct at the third stage when I recall the previous comment I obtained. (P11)”

Second, offering too many suggestions may decrease the willingness of solving difficult problems at the end of the process. For example, people gave up solving high-level issues at the last stage of the editing process. They felt that when they solve the content-level issues, all their previous efforts made to fix language-level issues were in vain.

“In the last stage, the feedback suggests that adding one connecting sentence in a specific paragraph increase the unity of the paragraph. I feel that all my previous efforts would be in vain if I fixed this issue. (P12)”

4.6 Discussion and Implications

In this section, we reflect on the interesting findings, the major challenge of the feedback presentation strategies explored, and propose suggestions of online feedback system design for writing support.

4.6.1 Feedback categorization guides learning behaviors

The results showed that the feedback categorization helped people structure feedback of different types and benefits people, ranging from identifying weaknesses to increasing

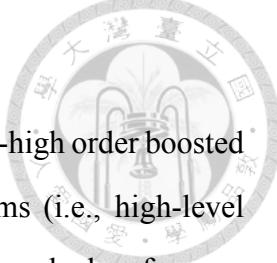
awareness to promoting learning behaviors. The idea of feedback categorization is inspired by writing rubrics, which are an assessment tool used to evaluate task performance over a range of criteria. In this work, we used an analytic rubric, which uses discrete criteria to assess a specific aspect of the work, to classify writing feedback. An analytic rubric has been proven to help people improve learning and performance [35, 59]. It makes assessment criteria transparent and people can use it to self-assess their performance; moreover, it also encourages reflective practice [49]. We have seen similar positive effects on our selective category structure followed by the best-known writing rubric [34].

Previous studies have used structure (e.g., rubrics or guiding questions) to guide crowd workers lacking the expertise to produce more specific and useful feedback [72, 50, 27]. By contrast, this work focuses on the use of a structure to support writers to interpret and integrate feedback. To achieve high-quality outcome, we suggest that a well-designed structure should be used to support not only feedback providers but also receivers for reducing knowledge gap between feedback generation and utilization.

4.6.2 Flexible support for varying preference

People require varying writing supports according to their writing abilities, current writing stage (drafting, early stage of revision, or editing), or preference. In the iterative writing process, an intelligent feedback system has to keep flexibility to adjust several important factors including the total number of comments, the ratio of high-level and low-level feedback, and ways of feedback presentation. In this work, we design a useful tool to present expert feedback in a structured way and explore the influence of three types of feedback strategies. The findings indicated the strength and weakness of different strategies and provided a guideline for further feedback system design. Moreover, our tool enables possibilities to rapidly explore the complex connection between feedback and writing performance.

4.6.3 Low-to-high sequence facilitates difficult problem solving



The most interesting finding is that people obtained feedback in low-to-high order boosted self-confidence and developed momentum to solve difficult problems (i.e., high-level writing issues). At the beginning of the editing process, when people may be less focused or may not be familiar with the content of the writing, low-level feedback can navigate them to read through the whole articles and solve low-level language issues simultaneously. Through the completion of a series of micro tasks, people become more aware of meaning, boost confidence, and gain a sense of momentum to move forward and overcome the more difficult content or organization issues. This finding is related to the previous crowdsourcing literature [9]. *Cai et. al.* indicated that performing low-complexity tasks (e.g. fix spelling or punctuation error) on the same operation helps develop momentum to keep doing. In addition, people can easily transit low-complexity tasks to high-complexity tasks (e.g. paragraphing) on the same content. However, we are tackling more difficult problems about dynamic task chains and more complex writing tasks.

The goal of this work is to examine how feedback presentation strategies affect revision behaviors and which strategies lead to better revision performance. In this situation, the revision activities may or may not be triggered by feedback, and the operation may not be the same as the comment suggested. Thus, the task chain is dynamic. Moreover, the high-level task of this work is about fixing real content or organization issues which are more diverse and complicated than paragraphing or shorten tasks, which are tested in *Cai et. al.*'s work. However, the findings obtained from *Cai et. al.*'s work also provide a great fundamental understanding of complex revision behaviors.

In this work, the findings suggest that using the low-to-high editing workflow can guide people to solve content or structure issues. However, applying this strategy should carefully consider the number of low-level feedback because people who spend much effort on low-level issues may diminish the willingness of performing large-scale revision in the end.

4.6.4 Decrease the workload of low-level, repetitive tasks

In this experiment, we observed that people tended to make no change at the last stage in the sequential workflow. After contributing a lot of effort on fixing a large number of low-level issues, people had lost interest in handling with high-level feedback at the end of the process. There are two possible reasons for this situation. First, a large number of repetitive and monotonous tasks may result in boredom and decrease work performance [20]. Second, people tend to give up making large-scale revisions due to a feeling of loss aversion, which is a phenomenon that people prefer avoiding losses to acquiring gains in decision-making under risk [36].

To reduce the negative effect of overwhelming low-level feedback, automated solutions can be applied to aggregate similar and common issues and provide an aggregated diagnostic feedback for highlighting common issues. Therefore, we suggest to provide a convenient way for writers to correct the common errors.

Moreover, according to the finding of momentum development, we suggest that distributing less low-level feedback around the content containing high-level issues may be a possible way to guide people to solve high-level problems with no obstacles.

4.6.5 Resolving editing conflicts and mental obstacles

In the sequential workflow, separating related feedback into different stages may result in misunderstanding or confusion. The two or more comments located in the same sentence may provide feedback from different perspectives; therefore, they have to be considered simultaneously rather than separately. In addition, due to the worry that high-level issues may overwrite low-level issues, high-to-low process prevents people from implementing high-level feedback at the beginning of the process.

To resolve spatial conflicts, we suggest grouping related comments and presenting them at the same stage. To avoid mental obstacles, we suggest increasing transparency of the whole process by providing an overview, and then people can foresee possible conflicts and reduce the mental barriers resulting from fear of the unknown.

4.7 Limitations and Future work



While our system was evaluated by a limited number of participants, this work contributes the first writing support system that enables flexible revision workflows for supporting effective revision behaviors. In addition, while there is no conclusive relationship between different feedback strategies in the field experiment, such finding is consistent to previous English writing tutors' suggestions [60, 3, 57] about high-to-low or low-to-high, but this work contributes the empirical understandings of how three types of feedback strategies (ALL, HML, LMH) influences editing behaviors.

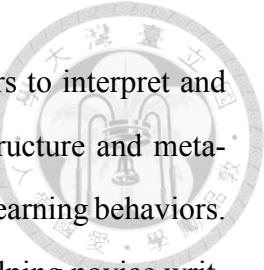
This work is attempt to explore an important but difficult research question about how does feedback facilitate high-quality revision results. This is still a open question while previous studies have had inconclusive findings. For example, Underwood et al. compiled a large of studies and concluded that more research is needed to determine the relationship between level of feedback and writing quality and revision practices at different times for different purposes [68]. However, our system ReviseO offers great opportunities to boost further research explorations of this critical question.

In addition, some usability issues may decrease the overall performance. For example, many participants complained of a linking problem between the comment and its error location. In the current system, we only provided one-way linking from the error location to the comment but did not implement the reverse linking. In the further improvement, we will enhance the user experience of the interface.

Furthermore, according to the results, we observed that ReviseO may benefit writers having basic writing abilities for self-learning. The novice writers are required to obtain basic writing knowledge before using this system. In the future, we will provide basic writing training and scaffolding writing structure for facilitating novice learners to acquire basic writing ability.

4.8 Conclusion

In this paper, we explored how structured feedback can support writers to interpret and integrate feedback into revisions in a structured way. A rhetorical structure and meta-feedback can help writers reflect on their own weaknesses and promote learning behaviors. Structured feedback can also enable flexible revision workflows for helping novice writers reflect on their behaviors and develop good strategies. Insights from the study suggest that designing an effective feedback framework has to support reflection and awareness in a flexible revision process. With further improvement, we believe that Feedback Orchestration can enable collaboration between writers and feedback providers for supporting novice writers to improve their work.







Chapter 5

How Feedback Affects Revision Quality?

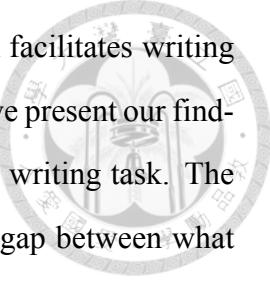
5.1 Introduction

External feedback helps people not only learn conceptual knowledge but also improve the quality of creative work in the iterative process. Previous studies have shown that seeking diverse feedback from crowds helps improve early-stage works in many domains, especially in design (e.g., [51]). However, few studies investigate whether crowd can contribute feedback that is more useful than an expert, and in what ways.

To better understand difference of feedback generated from experts and crowds, we first conducted a content analysis on a collection of feedbacks collected from writing experts and general crowd workers. From the results of content coding, crowd workers tended to generate low-level feedback that focus on local issues (e.g., spelling and grammar) rather than high-level feedback that focus on structures (e.g., topics and organization) and an individual worker contributed only a specific type of feedback when there is no guidance. By contrast, an expert could contribute more diverse types of feedback in a consistent way. However, high-level feedback were also lacking.

A field experiment was conducted to understand how ESL (English as a Second Language) learners revise their writings based on three sources of feedback, including expert

feedback, generic crowd feedback and StructFeed, in an iterative writing-revision process. The goal of this work attempts to explore whether ensemble feedback facilitates writing revision and improve output performance. In this late breaking paper, we present our findings from post-task interviews with the ESL learners participating the writing task. The results suggest that ensemble feedback helps learners understand the gap between what they intend to convey and what readers interpret and facilitates deep reflection.



5.2 Crowd Feedback vs Expert Feedback

To understand how expert and crowd provide feedback, we conducted a pilot study which collected open-ended feedback for 6 essays from two sources. Three experts were recruited from Wordvice¹, a professional English proofreading and editing service. Six crowd workers were recruited from Amazon Mechanical Turk. Each essay was around 300 to 400 words. The cost of collecting expert and crowd feedback was \$16 and \$2 per essay, respectively.

We evaluated two sources of feedback from two perspectives (feedback type and feedback form). Feedback types are content, organization, vocabulary, language use (grammar and others), and mechanics, which follows a standard classification in writing literature[34]. Feedback forms (or strategies) are direct editing, general comment, and specific comment. Direct editing is direct feedback strategy; general/specific comment is indirect feedback strategy [add ref]. The specific comment indicates an error at a particular location of word, sentence, or paragraph; on the other hand, the general comment indicates the error indirectly.

In the study, we collected 316 expert feedbacks and 79 crowd feedbacks in total. The average individual results are shown in Figure 5.1. An expert generated a higher quantity and type of feedback than an individual crowd worker. However, collecting feedback from crowd spent less time and cost.

Furthermore, we compare two kinds of feedback from individual and overview perspective (see Figure 5.2 and 5.3). For individual perspective, different expert used a con-

¹<http://wordvice.com.tw/toefl-writing-editing/>

	Expert	Crowd
Time	24~48 hrs	10~30 mins
Quantity	53.67	13.17
Cost (USD)	\$16	\$2
Feedback Type	more	less
Feedback Form (Direct/Indirect)	2.7	3.2



Figure 5.1: The overall comparison between an expert and an individual crowd worker.

sist way to provide feedback, but crowd worker did not. Expert provided diverse types of feedback on each essay; crowd tended to focus on a specific type of feedback, so some types of feedback were missing on an essay. However, for overall perspective, crowd workers achieved more high diversity of feedback just as expert did.

Expert and crowd had similarity and dissimilarity on feedback provision. For similar perspective, both expert and crowd generated more low-level feedback (vocabulary, language use, and mechanics) than high-level feedback (content and organization). Low-level feedback was provided by direct editing and high-level feedback was provided by a comment. For dissimilar perspective, expert provided more vocabulary feedback than crowd. Expert directly demonstrated a better word choice and added a specific comment to reminder the writer. In addition, all crowd workers lacked organization feedback.

5.3 Experiment: Writing Revision Affected by Feedback of Different Types

To examine whether crowd-generated structural feedback helps ESL learners improve their writing, we conducted a within-subject, order-balanced experiment for understanding how different feedback-generation conditions influence writing revision. Three methods were compared: expert feedback, free-form crowd feedback, and structural crowd feedback. We asked each participant to write an essay on their own first, and invited them to

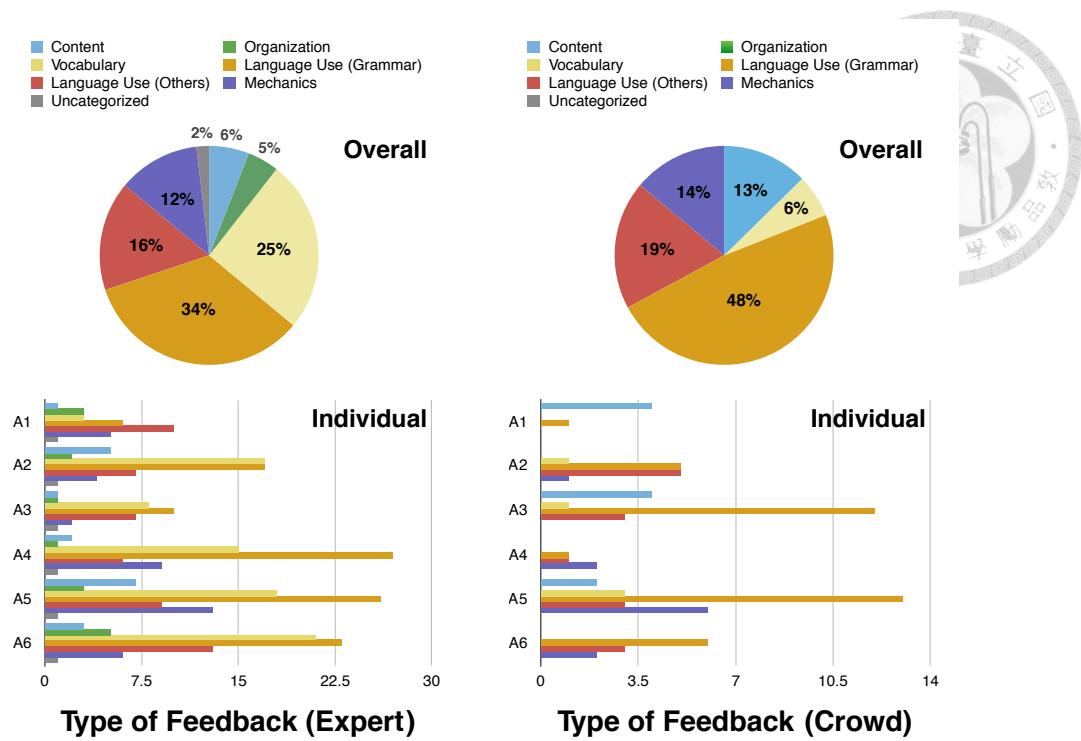


Figure 5.2: The detailed comparison of feedback type between expert and crowd (individual view and overall view).

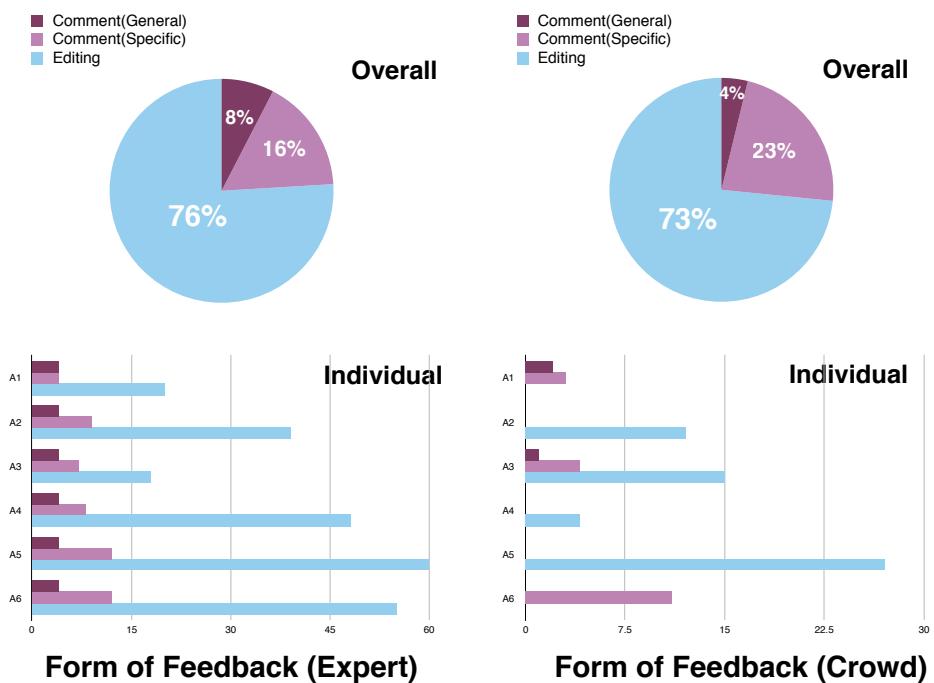
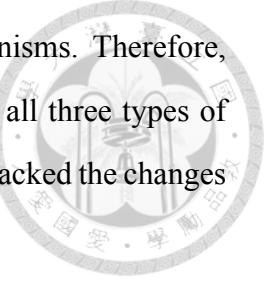


Figure 5.3: The detailed comparison of feedback form between expert and crowd (individual view and overall view).

revise their writing iteratively for three times. Each time they would receive feedback on their current version from one of the three feedback-generation mechanisms. Therefore, by revising the writing three times, the participants would experience all three types of feedback. The order of feedback conditions was counterbalanced. We tracked the changes they made to each revision this iterative writing process.



The first feedback-generation mechanism (C1) is traditional writing feedback that's generated by one expert in Wordvice², a professional online editing, and proofreading service. We used a particular service named TOEFL writing editing. The editor graded, edited, and provided diverse feedback on the structure, content, grammar, and word choices of the article based on the grading rubric of TOEFL iBT. The second condition (C2) is free-form writing feedback that's generated by a single crowd worker recruited from a crowdsourcing platform. The crowd worker is asked to provide writing suggestions about how to improve the unity and coherence of writing. The third condition (C3) is structural feedback with topic annotations, relevant keywords, and hints for writing generated by StructFeed.

5.3.1 Participants

We recruited 18 volunteers aged 19-35 years (56% male) through online community postings in Facebook and Bulletin Board System. Each participant is self-motivated to participate the experiment with a goal of practicing his/her writing skill.

5.3.2 Task and Procedure

The procedure of the experiment consisted of one writing stage and three rewriting stages. Each participant was required to complete one writing task and three rewriting tasks by using the three feedback conditions in a week. The time interval between any two tasks should be at least more than one day. The writing and rewriting tasks were enabled by using Google Docs. The experiment was held remotely with the help of an online video con-

²<http://wordvice.com>

ferencing system, appear.in³. The participants were asked to share their screen throughout the whole process.



Writing stage

In the writing stage, there were four main steps required to be completed: proficiency test, 5-minute writing training, writing task, and grammar checking. Firstly, each participant was required to test their English ability using the Cambridge English Proficiency Assessment⁴. Their test scores were normalized as a 1-to-5 index of English proficiency level. Secondly, they took a 5-minute writing training regarding what is paragraph structure and how to achieve paragraph unity. Next, they were asked to write an essay with a given topic in 30 minutes. The topic was randomly assigned from the three topics selected from a TOEFL writing topic collection. Lastly, they were allowed to check grammatical errors of their writing with Grammarly⁵, a popular online grammar checker.

Rewriting stage

In the rewriting stage, participants would see the previous version of writing with writing feedback generated from one of three feedback-generation methods. They were asked to revise the article based on the feedback they received, and they were free to change anything they wanted in 30 minutes. After the rewriting task, they were also allowed to use Grammarly to correct the spelling or grammatical issues. Participants were invited to participate a 15-minute interview after the last rewriting stage.

5.3.3 Measures

Expert Evaluation

To evaluate the quality of writing, we recruited two experts with more than five years of ESL teaching and training experience to rate all essays independently blind to the condition

³<https://appear.in>

⁴<http://www.cambridgeenglish.org/test-your-english/adult-learners/>

⁵<https://www.grammarly.com>

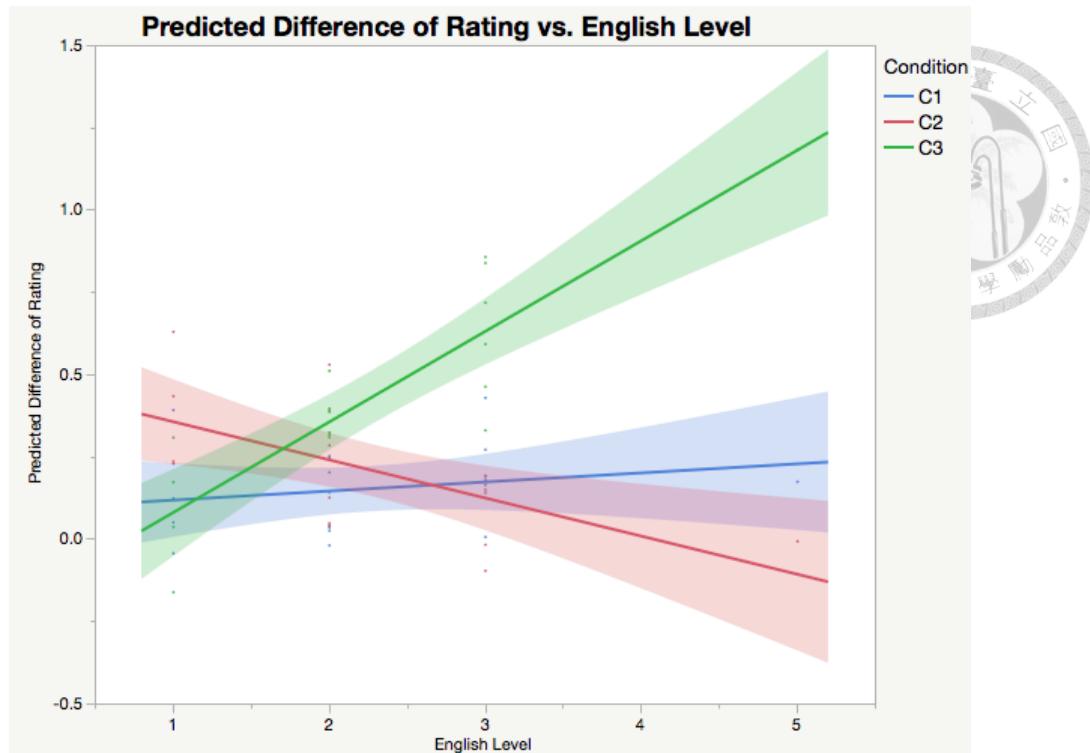


Figure 5.4: Relation between Difference of Rating, English Proficiency Level and Feedback condition. Points shown are results predicted by the linear mixed-model but not the raw data. Fitting lines are added to illustrate the trends.

based on the writing scoring rubrics of TOEFL iBT. The rating scale is from 0 to 5 with a 0.5 interval. The higher the rating, the better is the quality of writing.

Two raters independently rated all essays blind to condition. Ratings from the two raters were highly correlated, showing sufficient inter-rater reliability (Pearson $r = 0.93$, $p < .0001$). For each writing, ratings from the two raters were averaged to represent the quality of that writing.

Difference of Rating

To evaluate whether revision improves the quality of writing, we calculate the difference of pre-revision rating and post-revision rating for each revision by subtracting the pre-rating score from the post-rating score of a rewrite.

5.3.4 Statistical Results

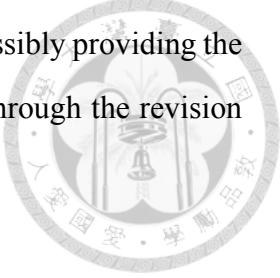
How Different Feedback-Generation Methods Influence Revision?

The study was a within-subject experiment, in which each participant experienced all three types of feedback-generation method. To account for the interdependency among the ratings of different versions of writing as they were all authored by the same individual, we used mixed-model ANOVAs to analyze the data. In the linear mixed model, individual participants were modeled as a random variable. Topic of writing, trial of revision, prior English proficiency level (measured by Cambridge English Proficiency test), and pre-revision rating of the rewriting was included in the analysis as controlled variables. Pre-rating of the writing is included as how well a writing is at the beginning may further influence the magnitude of possible improvement on the writing later. Feedback condition and its interaction with individual's English proficiency were the key independent variables. We built the statistical model and conducted the analyses using SAS' JMP 10.

From the mixed-model ANOVA, there was a marginal main effect of feedback condition on the difference of rating, $F[2, 30.06] = 3.0, p = .06$. Post-hoc contrast analyses further show that learners better improved their writing when feedback was provided by StructFeed (C3) than by expert (C1), $F[1, 30.41] = 5.31, p < .05$. This provides initial evidence that StructFeed can help learners' practice and improve their writings.

What's surprising and really interesting is that we found a significant interaction effect between feedback condition and English proficiency, $F[2, 30.01] = 6.45, p < .005$. As Figure 5.4 shows, whether feedback mechanisms affect difference of rating would depend on one's English level. C1 (expert feedback) and C2 (free-form crowd feedback) were shown to have limited utility on improving writing, while C3 (StructFeed-generated feedback) was shown to improve the quality of writing when the learner is with high English proficiency (e.g., students with English proficiency of 5). The pattern is understandable as beginners of writing tend to focus on shallower writing problems, such as spelling and grammatical issues. The high-level, structure-oriented feedback may be seen as being unuseful to this the beginners. On the other hand, advanced learners with higher language proficiency are likely to know already what are the shallow writing issues even when

there's no feedback. What's needed to them would then be feedback that focuses on organizational, structural aspects of writing. Therefore, StructFeed was possibly providing the right type of feedback in-time, which help to improve their writing through the revision process.



5.3.5 Insights from Interview Data

StructFeed Helps Identify the Gap between Intention and Interpretation

StructFeed aggregates multiple readers' annotations on topic sentence and relevant keywords in an essay. Those weighted annotations help writers understand the gap between their intentions and readers' interpretations.

“StructFeed helps me understand inappreciable problems. I'm so surprised the difference between my intention and the readers' interpretations. It makes me reflect whether my writing is not clear.” (P7)

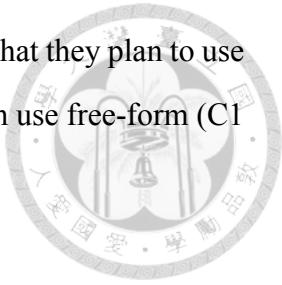
“I think that StructFeed provides the information that I cannot obtain through self-learning; on the contrary, I can learn by myself for solving most issues that C1 and C2 focus on.” (P3)

Most participants found usefulness for obtaining StructFeed feedback; however, one participant (P8) did not. She reported that she found difficulties to improve the writing, even if she understood the precise location irrelevant sentences. She preferred direct editing rather than indirect suggestions. We found that the usefulness of StructFeed might be influenced by the English ability of individual.

Combined Usage of StructFeed and Free-Form Feedback

StructFeed provides a clear goal and precise location of error sentences; on the other hand, free-form feedback (C1 and C2) allows reviewers to directly edit the writing aligning with detailed explanations to the writing issues. 8 participants (P3, P5, P7, P9, P10, P11, P12, P15) mentioned that they want to use StructFeed for their daily writing practice, 6

participants (P5, P11, P12, P7, P9, P15) of them indicated that they would like to combine StructFeed with free-form feedback(C1 or C2). They also mentioned that they plan to use StructFeed for multiple iterations in the early stage of writing and then use free-form (C1 or C2) lately.



StructFeed Promotes Self-Reflection

StructFeed allows users to perceive the gap between their intentions and other readers' interpretations. The gap promotes users to reflect and identify their deficiency, resulting in increasing motivation to improve a specific ability.

P7 said, "In my essay, I used an example of Michelin Three-star Restaurant to describe an expensive restaurant. I'm very surprised that no one annotates it as relevant keywords. I originally think that everyone should know "Michelin Three-star" is related to high price and expensive, but the fact that it is not a common example for other people. Therefore, I will carefully choose an precise and understandable example to describe my idea next time."

P7 said, "I found that many readers had mis-annotated concluding sentence as a topic sentence. It makes me think that my writing ability to paraphrase is not enough and I want to enhance it."

Instability of Crowd Feedback

Free-form feedback obtained from crowd workers has unstable quality. Some participants (P7, P10, P12, P13, P15) indicated that this type of feedback was few, and the content was too general and ambiguous; in contrast, some participants (P5, P10) found it helpful and the explanation was clear.

"I don't understand what he means. His comments containing terminology and it's hard for me to capture the key points." (P15) or "Few and general! He only mentioned that the paragraph is poor without any explanation and concise steps. I don't know how to modify it." (P12) (Negative)

“The comments focus on word choice and the explanation is good.” (P10)

or “The comments were effective. He indicate that my topic sentence is not clear.” (P5) (Positive)



5.4 Discussion

5.4.1 The Cost and Benefit of StructFeed

At the system building side, results showed that ML-based method could outperform the crowd-based method regarding identifying topic sentence and relevant sentence. However, the crowd-based method achieved the best performance regarding identifying irrelevant sentence. Compared to the ML-method, StructFeed not only identifies and locates writing issues but also provides diverse perspectives of how a diverse pool of potential readers (i.e., crowd workers) may interpret the writing. For example, participant P7 reported that StructFeed helped her realize that the example she used in her essay might confuse other people. With the type of feedback from StructFeed, she said she would choose a more suitable example in her further revision. It's also clear that StructFeed is more flexible than ML-based method because it can be applied to support writing of different topics and genres without the needs of training, the availability of corpus or prior knowledge for composing decision rules.

5.4.2 The Utility of StructFeed Depends on Learners' Level of Proficiency

From the field testing of using different feedback generation methods to assist ESL learners' writing revision, we found that while StructFeed in general helps revision and improves the quality of writing, structural feedback is more useful to learners when they're with higher levels of English proficiency. The finding is understandable as beginners with low English proficiency may tend to focus on shallower issues of writing, such as word spelling and grammatical correctness. Those high-level, structure-oriented feedback may

not be very useful to the beginners presumably due to their limited linguistic knowledge and limited experiences in English writing. On the other hand, advanced learners with higher language proficiency may already know what are the shallow writing issues and how to handle them. They're also likely to have less problems in aspects like spelling and simple sentence composition. What's new and useful to them would then be feedback that focus on organizational, structural aspects of writing, which explains why StructFeed is especially useful to overcome advanced learners' writing barrier.

One implication to future design is the need to consider combining different methods of feedback generation to offer writing feedback that's sensitive and adaptive with respect to learners' levels of writing. Our user study also shows plausibility of this approach in terms of usability as participants of the study essentially experienced three different feedback conditions throughout the iterative revision process, and this did not seem to cause confusion. In the future, we can further consider how to better integrate crowd, expert and natural language processing techniques to provide support for different types and levels of writing issues.

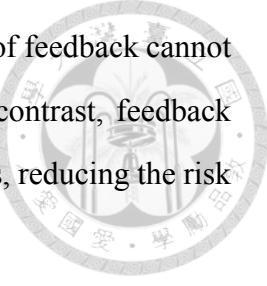
5.4.3 Gap between Expert Reviewer and Novice Writer

Reducing the knowledge gap between expert reviewer and novice writer is required for a feedback system. According to the experiment result, expert feedback conditions got smaller difference between pre-rating and post-rating, comparing with the StructFeed condition. That is, the improvement of expert feedback condition is smaller than the StructFeed condition. In addition, in the interview, many learners reported that they were eager to communicate with the expert reviewer for clarifying the meaning of feedback. Therefore, building shared knowledge or enabling communication between reviewer and writer is required in a feedback system.

5.4.4 Macro-Task vs. Micro-Task

The quality of feedback generated by mechanisms that leverage micro-tasks (StructFeed) is shown to be more stable than feedback generated with macro-tasks (expert feedback and

free form crowd worker feedback). For the general crowd feedback condition, a single crowd worker was recruited to generate writing feedback. The quality of feedback cannot be guaranteed to be useful based on varying ability of workers. In contrast, feedback generated by StructFeed aggregated the answers from multiple workers, reducing the risk and uncertainty of obtaining low-quality results.







Chapter 6

Reflection After/Before Practice

6.1 Introduction

Developing professional drawing skills requires not only time and effort but also deliberate practice based on immediate, effective feedback [17]. However, obtaining effective drawing feedback is challenging. First, drawing feedback is usually generated by experienced instructors and only available in school or in a small physical art studio. Second, instructors usually provide instant personalized visual feedback directly on the canvas or demonstrate the whole creative or revision process, along with detailed explanations. Such high-quality personalized feedback can be obtained through one-on-one tutoring, but it is costly and has a limited pool of qualified experts.

Many interactive drawing systems have been developed to support people to draw more accurately by providing automated corrective feedback, guidance, and tutorials [15, 33, 19]. The corrective feedback allows people to be aware of “knowing” the gap between the goal and the current status. Step-by-step instructions or tutorials guide people to follow the procedural steps to recreate a reference image as accurately as possible. However, good drawing support has to enable people to understand high-level conceptual knowledge rather than to memorize the low-level steps.

This work attempts to apply learnersourcing to generate personalized drawing hints for supporting future learners to develop drawing knowledge, useful skills and diverse coping strategies. Prior studies have successfully used learnersourcing to enable people

to collectively generate useful annotations or hints while engaging in the learning process [71, 26]. Glassman et al. presented two workflows that allowed learners to generate personal hints based on what problems they had recently solved for engineering circuit design. Moving beyond learnersourcing personal hints for seeking optimized solutions, we tackle an open-ended, ill-defined problem of seeking “how good the answer is” instead of simply “whether the answer is true”. In addition, we collect personal drawing hints to support learners with diverse ability levels and needs.

In this work, we explore how learnersourcing can be applied to support drawing learning. First, we introduce the concept of learnersourcing drawing support. Next, we present ShareSketch, a web-based drawing system that enables learners to practice drawing, review the drawing process, and share their works with others. Moreover, we have designed a reflection workflow that allows learners to contribute annotations about what they learned and why they consider it helpful, interesting, or difficult while watching from other’s creative process. An annotation, indicated as a learning point, consists of a clip extracted from the process, description, and explanation. Those personal learning points can benefit future learners while experiencing similar situations. More importantly, we propose *before/after-practice reflection workflow*, which is an extended reflection workflow along with each short practice. People can quickly practice lessons learned from the before-practice workflow; they also can enhance or revise their findings in the after-practice workflow.

In the pilot study, we evaluated our before/after-practice reflection workflow with eight participants recruited from Facebook. They were asked to perform two reflection tasks before and after a short practice and we compared both annotations. We found that before-practice annotations are higher level than after-practice annotations. Furthermore, participants applied their before-practice learning points as subgoals; they revised or extended more detail their findings in after-practice annotations when facing conflicts on their actual performance.

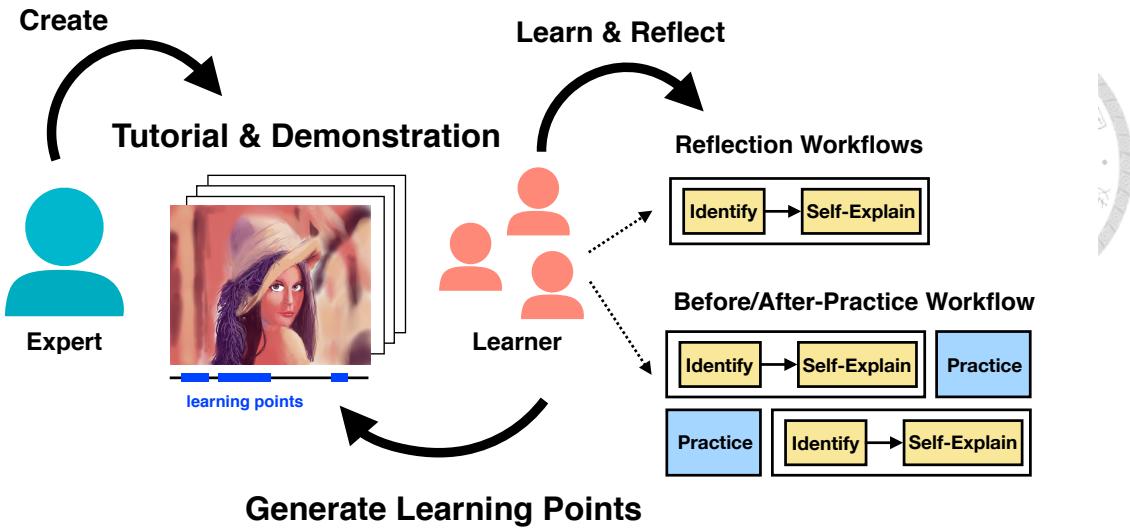


Figure 6.1: The overall of learnersourcing drawing support.

6.2 Related Work

Prior studies have developed interactive drawing assistants to support novice learners to develop drawing skills by providing automated corrective feedback, guidance, and tutorials [15, 33, 19]. They used sketch recognition [15] or automated extracting techniques [33] to generate visual guidance for helping people draw as accurate as a reference image. Through direct guidance or step-by-step instructions, learners can achieve better results after iterative modifications.

In this work, we focus on collecting personal learning points including high-level concept, low-level details, and personalized strategies while a learner watches other's drawing process. Through our workflow design, learners can both contribute useful information and enhance drawing skills by reflection and practice.

6.3 Learnersourcing for Drawing Support

We propose a framework that allows people to enhance their learning by practice and reflection, and then contribute learning points for supporting future learning (Figure 6.1). In the framework, people can draw and review the process for self-reflection on an online drawing system. Other people can contribute their learning points while practicing

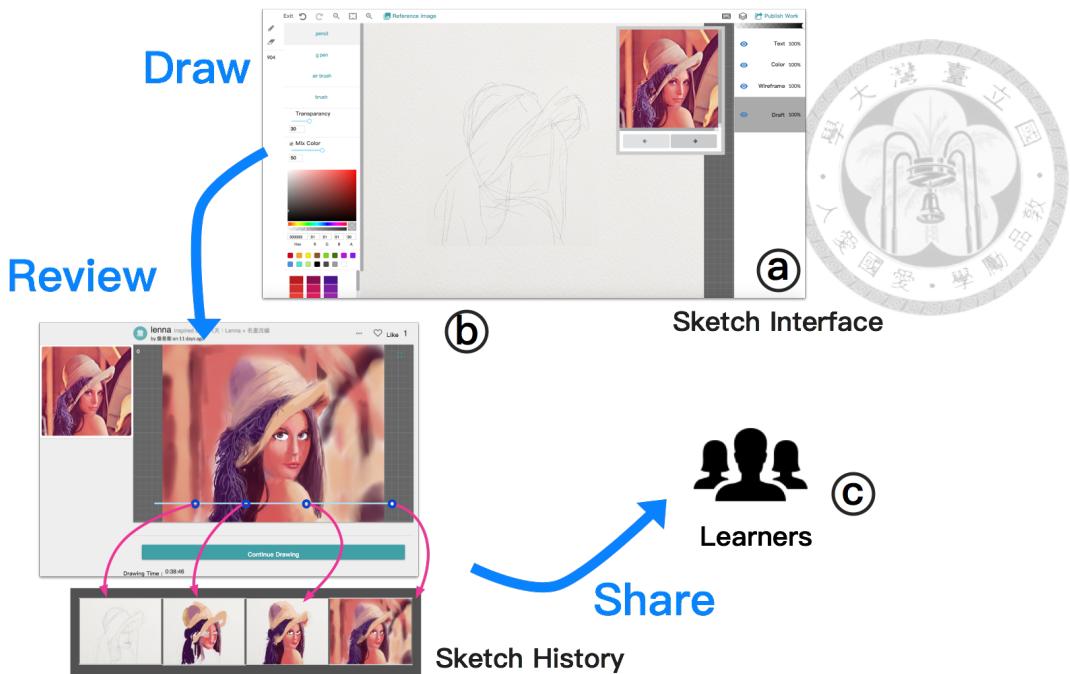


Figure 6.2: ShareSketch augments web-based drawing system with an interactive timeline. A user can create a drawing by a sketch interface (a), review the drawing process by an interactive timeline interface (b), and share the process to others (c).

and watching a drawing process in our reflection workflows. Both drawing authors and viewers can benefit from reflection and practice.

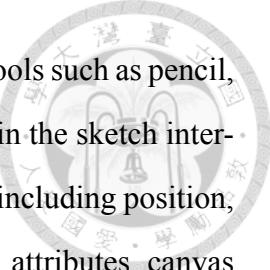
In this work, we will introduce our online drawing system called ShareSketch and reflection workflows for enabling learners to generate useful learning points.

6.4 ShareSketch: Draw, Review and Share

ShareSketch is a web-based drawing system built with Javascript and WebGL that allows a user to create a drawing, review the creative process and share with other people. The goal is to enable people to review their drawing process for self-reflection and promote social learning. The system consists of two main components: a sketch interface, and an interactive timeline interface (Figure 6.2).

6.4.1 Sketch Interface

ShareSketch supports WACOM tablet and provides common drawing tools such as pencil, brushes, or an eraser. Users can also draw amongst multiple canvases in the sketch interface (Figure 6.2). The interface can record detailed drawing behaviors including position, speed and pressure of strokes, and tool usage behaviors including tool attributes, canvas transforming and undo commands. Those recorded behaviors will be used to extract some patterns for detecting abnormal events.



6.4.2 Timeline Interface for Sketch History

The interactive timeline interface is designed to allow a user to review the history of drawing. Users can replay the whole creative process or go back to any time point to examine the work-in-progress. It also allows people to add an annotation on a clip which is selected from the process.

6.5 Before/After-Practice Reflection Workflow

To enable people to generate learning points from watching other's creative process, we designed a two-stage reflection workflow including *Identify* and *Explain* stage. On the *Identify* stage, people can identify a clip from the process by selecting a start and an end point, and then describe what they learned from the selected clip (Figure 6.3). After three iterations, three annotations will be passed to the *Explain* stage. In this stage, people can choose the most helpful one from the three annotations and explain why they think it is helpful to them.

By using this workflow, we can collect three types of annotations from learners, including helpful, interesting and difficult.

Furthermore, we design a before/after-practice reflection workflow that allows learners to generate learning points before or after a quick practice. Through practice, learners can self-assess the actual level of abilities and enhance the learning effect. In the before-practice workflow, learners can reflect on their own experience and apply learning points

Q1-1: Select the process that you learned something

Example :
start at: 25
end at: 77
Describe what do you learn in this process: I have learned how to construct the whole

start at: 964

end at: 1587

Describe what do you learn in this process:

Next

Figure 6.3: A learner performs a reflection task by identifying a clip and describe what it is, and then explain why they learned from the clip.

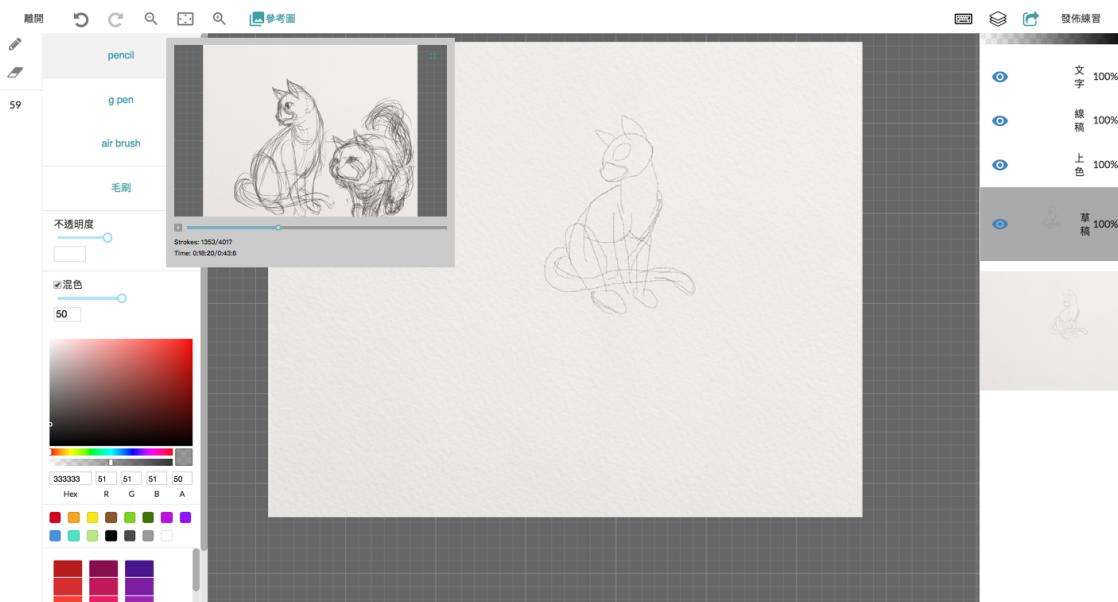


Figure 6.4: Learners are allowed to practice drawing based on other's creation process in a short practice task.

in the short practice. Moreover, in the after-reflection workflow, learners can re-assess their current status after a short practice and provide their revised or extended findings.

In a short practice task, learners can apply their learning points, and develop better coping strategies by tackling the actual obstacles.



6.6 Pilot Study

To better understand how learners identify learning points by watching others' drawing processes, we ran a pilot study that recruited online drawing learners to perform two reflection tasks, before and after a quick practice. We recruited 8 participants from Facebook. They were 5 male and 3 female self-motivated learners, 20-23 years of age, with only one design-related major. Each person was allowed to watch one of two drawing creative processes created by two experts and complete each reflection task before and after a short practice. The short practice task should be performed in approximately 5 minutes. They can use either a mouse or a drawing tablet to perform the practice task. The total time spent was approximately 30 minutes.

In the study, each participant was asked to identify three learning points, one interesting point, and one difficult point. Finally, each participant generated five before-practice and five after-practice annotations. We compared two types of annotations and summarized our observations in the following section.

6.7 Results and Findings

We totally collected 80 annotations, and of those, only one was missing an explanation. The results of learners' practices are shown in Figure 6.5 and the findings are presented as follows.

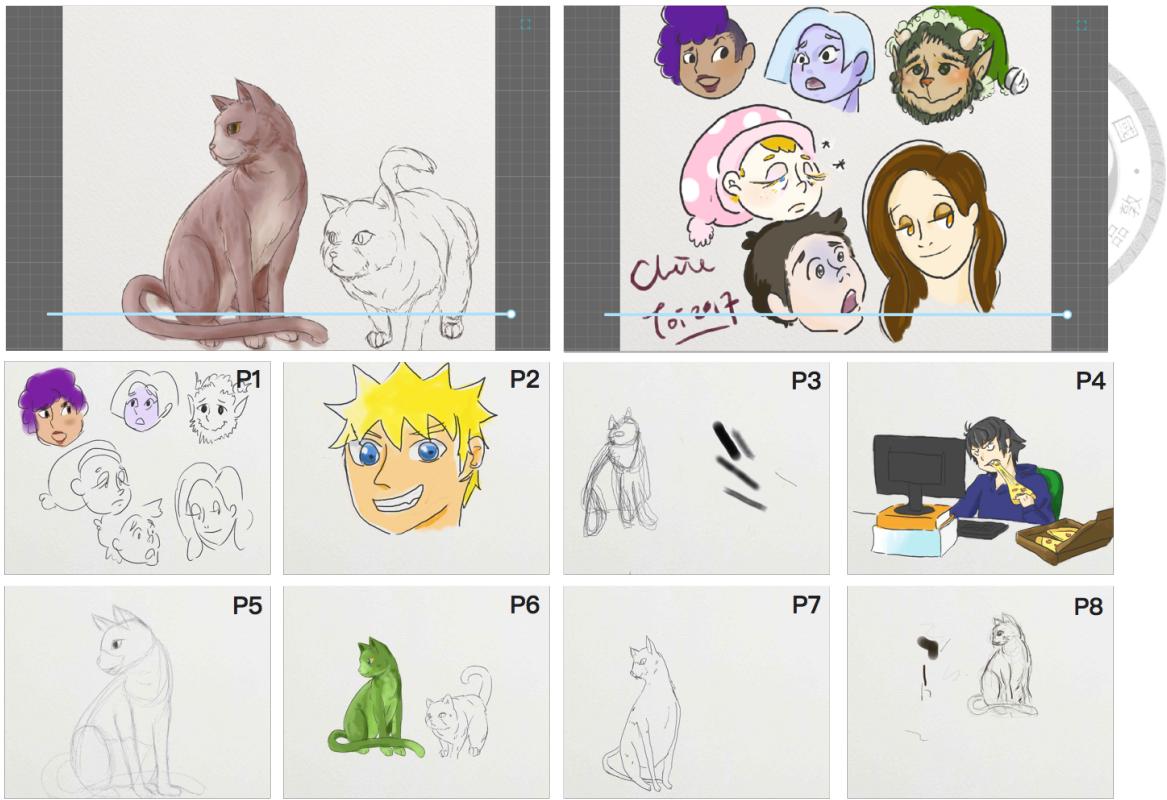


Figure 6.5: The above two images are two creative processes created by two experts. The below eight images are the results of participants' practices.

6.7.1 After-practice annotation augments before-practice annotation

We found that before-practice annotations are higher level than after-practice ones. In the before-practice annotations, participants identified a longer clip and described high-level concepts, including sub-goals and a general structure of procedural steps. In addition, participants described more details in after-practice annotations.

“I have learned how to construct a cat, from outlining to drawing the details.”

(P3, before-practice)

“I have learned how to define the direction of a human face by drawing a cross-axis to the face.”

(P4, after-practice)

“I have learned how to highlight the bright side of an object by using an easier to remove the color.”

(P4, after-practice)

6.7.2 Before-practice reflection vs After-practice reflection

Participants identified new learning points based on the difference between past experiences and others' drawing processes in the before-practice reflection task. On the other hand, they changed their perspective and revise their findings based on comparisons or conflicts between original thinkings and practicing results. In addition, two participants added the details or extend the findings in the after-practice task.

"I have never learned the impasto painting technique before. It looks so cool while adding one color layer by layer to construct the three-dimensional effect." (P5, before-practice, discover new thing)

"I found the difficulty of combining the multiple colors. It's not as easier as I first think. "

(P8, after-practice, change perspective)

"I learned how to create a human face with different emotion expression" (P4, before-practice);

"I have learned how to highlight the bright side of a face by using an easier to remove the color."

(P4, after-practice, add the details)

6.8 Discussion

6.8.1 Provide scaffolding for reflection and practice

Most participants created a drawing with low completeness during a short period of time; in addition, they reported the difficulty of practicing a specific skill in the later drawing process because they lacked domain knowledge and related skills to reach the goal. In the further improvement, we will design a scaffolding mechanism that supports people to practice learning points at any stage of drawing process.

Moreover, many participants used distinct words to describe similar concepts or situations. To effectively aggregate learning points, we plan to guide people to generate consistent content using a structured interface or a rubric [50].



6.8.2 Learning points as feedback enhances creative task learning

Feedback is the key component to help people improve the creative process and iterate toward better results. In this work, learning points serve as feedback benefits drawing creators and learners. Creators can increase confidence or motivation by obtaining others' feedback; learners can reflect what they learn and benefit to other learners.

6.9 Future Work

We will keep improving our system and workflow design. In addition, we will explore ways to generate useful learning points and facilitate effective practice. In the end, we envision our framework can contribute to online feedback system design and support other creative tasks.



Chapter 7

Conclusion

This dissertation introduced a never-ending creative learning framework in which users improve their work and develop their professional skills based on structural feedback and reflection revision workflow. We have presented three systems for generating effective feedback, facilitating revision, and supporting learning for creative tasks. This final chapter restates the contributions of this dissertation and discusses future directions.

7.1 Restatement of Contributions

To sum up, this dissertation has demonstrated the power of crowd and machine to support creative task solving and learning. The contributions of this work can be summarized as follows:

- **Crowdsourcing:** novel workflows for extracting semantic information from writing and collecting learning points from learners; structure patterns as scaffolding mechanisms that are designed to guide people to generate high-quality annotations, make sense of feedback, and effectively integrate feedback into revision.
- **Online feedback exchange:** an novel iterative feedback framework for supporting creative task learning; novel systems that support feedback generation process, feedback-driven revision process, as well as creative learning process.

- **Creative tasks solving and learning:** design implications and insights that supports future researchers to explore other learning tasks for creative domains.



7.2 Future Directions

This work provided evidence how crowd-based technologies could assist novice people to accomplish creative tasks, develop professional skills, as well as enhance the learning experience. Based on these design experiences, we propose directions of future research in computational feedback design.

7.2.1 Hybrid Combination of Crowd and Machine

Human intelligence and machine intelligence both have their own advantages. Humans are good at making sense of content and extract high-level semantic meaning from texts, even from poor-written texts, yet they require incentives to contribute their efforts and only have a limited of attentions. On the other hand, machines are good at dealing with large amounts of monotonous and repetitive tasks, extracting pattern from large amounts of structured data, but they still require labeled data to train good predictive models and only support for a limited domain. In addition, it is still hard for machines to process unstructured information, rare patterns, and information that requires high-level of semantic interpretation [12]. However, we found that understand that novice writers performed better while receiving structured feedback from StructFeed instead of free-form feedback from both a single crowd worker or an expert. Results suggest that mixed structured feedback generated from machine rules and crowd annotation is good solution instead of pure human or machine feedback.

The structured feedback in StructFeed is highly relied on human annotations and simple rule-based algorithms. Can machine intelligence be integrated into the feedback generation process or feedback-driven revision process? Can certain types of feedback be generated by machine learning techniques? How to make use of both human and machine intelligence to balance cost, time, and performance? For example, we could utilize NLP or

AI techniques to automatically detect writing issues and diagnose users' expertise. Based on these information, future systems can decide which types of annotations are required to collect from crowd workers and generate personalized feedback for support user with different levels.



7.2.2 Creative Knowledge Construction for Innovative Applications

Feedback is useful for support people to accomplish creative tasks; however, valuable tacit knowledge may exist in the context of practice. Collecting interaction between feedback and work may benefit to future learners. Therefore, constructing tacit creative knowledge is of paramount importance for supporting such creative domain learning. Rich knowledge bases have been proven successful for enabling various useful applications. For example, Wikipedia, ImageNet, ConceptNet and Google Knowledge Graph have helped many researchers or companies to build intelligent models or systems for image recognition, machine translation, conversation agents, other areas. Therefore, it is critical to collect and construct creative knowledge to support innovative applications.

7.3 Summary

In this dissertation, we present a never-ending creative learning framework that support novice people to solve creative tasks and develop their professional skills. By leveraging the power of crowd and machine, our systems can generate effective structural feedback for supporting writing and facilitating effective and revision. Moreover, we integrate reflection into our system design and enhance learning experience in practice. In the end, this never-ending creative learning framework presents a promising vision for solving difficult problem by enabling collaborations between human and human, human and machine, as well as human, crowd, and machine.





Bibliography

- [1] M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, and K. Panovich. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, UIST '10, pages 313–322, New York, NY, USA, 2010. ACM.
- [2] M. S. Bernstein, D. Tan, G. Smith, M. Czerwinski, and E. Horvitz. Personalization via friendsourcing. *ACM Trans. Comput.-Hum. Interact.*, 17(2):6:1–6:28, May 2008.
- [3] S. Blau, J. Hall, and S. Sparks. Guilt-free tutoring: Rethinking how we tutor non-native english-speaking students. *The Writing Center Journal*, 23(1), 2002.
- [4] M. Boekaerts. Self-regulated learning: where we are today. *International Journal of Educational Research*, 31, 1999.
- [5] M. Boekaerts, P. R. Pintrich, and M. Zeidner, editors. *Handbook of Self-regulation*. Academic Press, 2000.
- [6] D. Boud. *Enhancing Learning Through Self Assessment*. Routledge, 1995.
- [7] J. Brooke. Sus: A "quick and dirty" usability scale. *Usability Evaluation in Industry*, 1996.
- [8] J. Burstein, M. Chodorow, and C. Leacock. Automated essay evaluation: The criterion online writing service. *AI Magazine*, 25(3):27–36, 2004.
- [9] C. J. Cai, S. T. Iqbal, and J. Teevan. Chain reactions: The impact of order on microtask chains. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3143–3154, New York, NY, USA, 2016. ACM.

- [10] L. B. Chilton, G. Little, D. Edge, D. S. Weld, and J. A. Landay. Cascade: Crowd-sourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 1999–2008, New York, NY, USA, 2013. ACM.
- [11] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović, and F. players. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–760, Aug 2010.
- [12] R. M. Dawes, D. Faust, and P. E. Meehl. Clinical versus actuarial judgment. *Science*, 243:pp. 1668–1674, 1989.
- [13] J. Dewey. *How We Think: A Restatement of the Relation of Reflective Thinking to the Educative Process*. D.C. Heath and Company, 1933.
- [14] S. Dikli. An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment*, 5(1), 2006.
- [15] D. Dixon, M. Prasad, and T. Hammond. icandraw: Using sketch recognition and corrective feedback to assist a user in drawing human faces. In *Proc. CHI 2010*, 2010.
- [16] S. Dow, A. Kulkarni, S. Klemmer, and B. Hartmann. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '12, pages 1013–1022, New York, NY, USA, 2012. ACM.
- [17] K. A. Ericsson, R. T. Krampe, and C. Tesch-romer. The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 1993.
- [18] L. Faigley and S. Witte. Analyzing revision. *College Composition and Communication*, 32(4):400–414, 1981.

- [19] J. Fernquist, T. Grossman, and G. Fitzmaurice. Sketch-sketch revolution: An engaging tutorial system for guided sketching and application learning. In *Proc. UIST 2011*, 2011.
- [20] C. D. Fisherl. Boredom at work: A neglected concept. *Human Relations*, 46(3):395–417, 1993.
- [21] J. Fitzgerald. Research on revision in writing. *Review of Educational Research*, 57(4):481–506, 1987.
- [22] L. Flower and J. R. Hayes. A cognitive process theory of writing. *College Composition and Communication*, 32(4), 1981.
- [23] L. Flower, J. R. Hayes, L. Carey, K. Schriver, and J. Stratman. Detection, diagnosis, and the strategies of revision. *College Composition and Communication*, 37(1):16–55, 1986.
- [24] E. Foong, S. P. Dow, B. P. Bailey, and E. M. Gerber. Online feedback exchange: A framework for understanding the socio-psychological factors. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 4454–4467, 2017.
- [25] E. Foong, D. Gergle, and E. M. Gerber. Novice and expert sensemaking of crowd-sourced design feedback. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):45:1–45:18, 2017.
- [26] E. L. Glassman, A. Lin, C. J. Cai, and R. C. Miller. Learnersourcing personalized hints. In *Proc. CSCW 2016*, 2016.
- [27] M. D. Greenberg, M. W. Easterday, and E. M. Gerber. Critiki: A scaffolded approach to gathering design feedback from paid crowdworkers. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, pages 235–244, 2015.

- [28] N. Greer, J. Teevan, and S. T. Iqbal. An introduction to technological support for writing. Technical report, Microsoft Research Tech Report MSR-TR-2016-001, 2016.
- [29] N. Hahn, J. Chang, J. E. Kim, and A. Kittur. The knowledge accelerator: Big picture thinking in small pieces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2258–2270, 2016.
- [30] C. M. Hicks, V. Pandey, C. A. Fraser, and S. Klemmer. Framing feedback: Choosing review environment features that support high quality peer assessment. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 458–469, 2016.
- [31] A. S. Horning and A. Becker. *Revision: History, Theory, and Practice*. Parlor Press, 2006.
- [32] J. Hui, A. Glenn, R. Jue, E. Gerber, and S. Dow. Using anonymity and communal efforts to improve quality of crowdsourced feedback. In *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing*, 2015.
- [33] E. Iarussi, A. Bousseau, and T. Tsandilas. The drawing assistant: Automated drawing guidance and feedback from photographs. In *Proc. UIST 2013*, 2013.
- [34] H. L. Jacobs, S. A. Zinkgraf, D. R. Wormuth, V. F. Hartfiel, and J. B. Hughey. *Testing ESL Composition: A Practical Approach*. Newbury House, 1981.
- [35] A. Jonsson and G. Svingby. The use of scoring rubrics: reliability, validity and educational consequences. *Educational Research Review*, 2(2):130–144, 2007.
- [36] D. Kahneman and A. Tversky. Prospect theory: An analysis of decisions under risk. *Econometrica*, 47(2):263–291, 1979.
- [37] R. B. Kaplan. Cultural thought patterns in inter-cultural education. *Language Learning*, 1966.

- [38] J. Kim, J. Cheng, and M. S. Bernstein. Ensemble: Exploring complementary strengths of leaders and crowds in creative collaboration. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2014.
- [39] J. Kim and A. Monroy-Hernandez. Storia: Summarizing social media content based on narrative theory using crowdsourcing. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 2016.
- [40] J. Kim, S. Sterman, A. A. B. Cohen, and M. S. Bernstein. Mechanical novel: Crowd-sourcing complex work through reflection and revision. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 233–245, 2017.
- [41] A. Kittur, B. Smus, S. Khamkar, and R. E. Kraut. Crowdforge: crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST ’11, pages 43–52, New York, NY, USA, 2011. ACM.
- [42] S. Konstantinidis. Computing the edit distance of a regular language. *Inf. Comput.*, 205(9):1307–1316, 2007.
- [43] M. Krause, T. Garncarz, J. Song, E. M. Gerber, B. P. Bailey, and S. P. Dow. Critique style guide: Improving crowdsourced design feedback with a natural language model. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI ’17, pages 4627–4639, 2017.
- [44] A. Kulkarni. Turkomatic : Automatic Recursive Task Design for Mechanical Turk. pages 1–6, 2011.
- [45] C. Kulkarni, K. P. Wei, H. Le, D. Chia, K. Papadopoulos, J. Cheng, D. Koller, and S. R. Klemmer. Peer and self assessment in massive online classes. *ACM Trans. Comput.-Hum. Interact.*, 20(6):33:1–33:31, 2013.
- [46] C. E. Kulkarni, M. S. Bernstein, and S. R. Klemmer. Peerstudio: Rapid peer feedback emphasizes revision and improves performance. In *Proceedings of the Second ACM*

Conference on Learning @ Scale, L@S '15, pages 75–84, New York, NY, USA, 2015. ACM.



- [47] I. Leki. *Understanding ESL Writers: A Guide for Teachers*. Portsmouth, 1992.
- [48] G. Little, L. B. Chilton, M. Goldman, and R. C. Miller. Exploring iterative and parallel human computation processes. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 68–76, 2010.
- [49] J. A. Luft. Rubrics: Design and use in science teacher education. *Journal of Science Teacher Education*, 10(2):107–121, 1999.
- [50] K. Luther, J.-L. Tolentino, W. Wu, A. Pavel, B. P. Bailey, M. Agrawala, B. Hartmann, and S. P. Dow. Structuring, aggregating, and evaluating crowdsourced design critique. In *Proc. CSCW 2015*, 2015.
- [51] X. Ma, L. Yu, J. L. Forlizzi, and S. P. Dow. Exiting the design studio: Leveraging online participants for early-stage design feedback. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 676–685, New York, NY, USA, 2015. ACM.
- [52] R. J. Marzano. *The Art and Science of Teaching: A Comprehensive Framework for Effective Teaching*. VA: ASCD, 2007.
- [53] M. Nebeling, A. To, A. Guo, A. A. de Freitas, J. Teevan, S. P. Dow, and J. P. Bigham. Wearwrite: Crowd-assisted writing from smartwatches. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 3834–3846, 2016.
- [54] D. T. Nguyen, T. Garncarz, F. Ng, L. A. Dabbish, and S. P. Dow. Fruitful feedback: Positive affective language and source anonymity improve critique reception and work outcomes. In *Proceedings of the 20th ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1024–1034, 2017.

- [55] D. J. Nicol and D. MacFarlane-Dick. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2):199–218, 2006.
- [56] J. Noronha, E. Hysen, H. Zhang, and K. Z. Gajos. Platemate: Crowdsourcing nutritional analysis from food photographs. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST ’11, pages 1–12, New York, NY, USA, 2011. ACM.
- [57] A. Oshima and A. Hogue. *Longman Academic Writing Series 4: Essays (5th Edition)*. Pearson Education ESL, 5th edition edition, 2013.
- [58] S. Perl. The composing processes of unskilled college writers. *Research in the Teaching of English*, 13(4):pp. 317–336, 1979.
- [59] Y. M. Reddy and H. Andrade. A review of rubric use in higher education. *Assessment and Evaluation in Higher Education*, 35(4):435–448, 2010.
- [60] T. J. Reigstad and D. A. McAndrew. *Training Tutors for Writing Center Conference*. ERIC and NCTE, Urbana, IL, 1984.
- [61] D. Retelny, S. Robaszkiewicz, A. To, W. S. Lasecki, J. Patel, N. Rahmati, T. Doshi, M. Valentine, and M. S. Bernstein. Expert crowdsourcing with flash teams. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST ’14, pages 75–85, New York, NY, USA, 2014. ACM.
- [62] S. Rose, D. Engel, N. Cramer, and W. Cowley. Automatic keyword extraction from individual documents. In M. W. Berry and J. Kogan, editors, *Text Mining. Applications and Theory*. John Wiley and Sons, Ltd, 2010.
- [63] D. R. Sadler. Formative assessment and the design of instructional systems. *Instructional Science*, 18(2):119–144, 1989.
- [64] D. A. Schön. *The Reflective Practitioner: How Professionals Think in Action*. New York: Basic Books, 1983.

[65] D. A. Schön. *Educating the Reflective Practitioner: Toward a New Design for Teaching and Learning in the Professions*. San Francisco: Jossey-Bass, 1987.

[66] N. Sommers. Revision strategies of student writers and experienced adult writers. *College Composition and Communication*, 31(4), 1980.

[67] J. Teevan, S. T. Iqbal, and C. von Veh. Supporting collaborative writing with micro-tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016.

[68] J. S. Underwood. Effective feedback: Guidelines for improving performance. In *Proceedings of the 8th International Conference on International Conference for the Learning Sciences - Volume 2*, pages 415–422, 2008.

[69] V. Verroios and M. S. Bernstein. Context trees: Crowdsourcing global understanding from local views. In *Proceedings of the Second AAAI Conference on Human Computation and Crowdsourcing (HCOMP-2014)*, 2014.

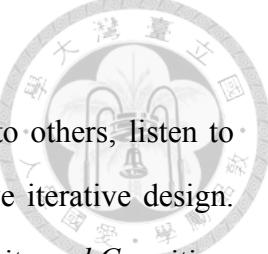
[70] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 319–326, New York, NY, USA, 2004. ACM.

[71] S. Weir, J. Kim, K. Z. Gajos, and R. C. Miller. Learnersourcing subgoal labels for how-to videos. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 405–416, New York, NY, USA, 2015. ACM.

[72] A. Xu, S.-W. Huang, and B. Bailey. Voyant: Generating structured feedback on visual designs using a crowd of non-experts. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 1433–1444, New York, NY, USA, 2014. ACM.

[73] Y.-C. G. Yen, S. P. Dow, E. Gerber, and B. P. Bailey. Social network, web forum, or task market?: Comparing different crowd genres for design feedback exchange.

In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, DIS '16, pages 773–784, New York, NY, USA, 2016. ACM.



- [74] Y.-C. G. Yen, S. P. Dow, E. Gerber, and B. P. Bailey. Listen to others, listen to yourself: Combining feedback review and reflection to improve iterative design. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*, C&C '17, pages 158–170, 2017.
- [75] A. Yuan, K. Luther, M. Krause, S. I. Vennix, S. P. Dow, and B. Hartmann. Almost an expert: The effects of rubrics and expertise on perceived value of crowdsourced design critiques. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 1005–1017, 2016.
- [76] H. Zhang, E. Law, R. Miller, K. Gajos, D. Parkes, and E. Horvitz. Human computation tasks with global constraints. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 217–226, New York, NY, USA, 2012. ACM.
- [77] B. J. Zimmerman. Self-regulated learning and academic achievement: An overview. *Educational Psychologist*, 25(1):3–17, 1990.