Enabling the collective to assist the individual: A self-organising systems approach to
social software and the creation of collaborative text signals

Andrew Francesco Chiarella

Department of Educational and Counselling Psychology
McGill University
Montreal, Quebec, Canada

Final Submission April, 2009

A thesis submitted to
Graduate and Postdoctoral Studies
in partial fulfilment of the requirements of
the degree of Doctor of Philosophy in Educational Psychology

© Andrew F. Chiarella, 2008
ACKNOWLEDGEMENTS

My sincere appreciation to Dr. Susanne Lajoie, for her constant encouragement. As well, your timely feedback was always fair and insightful. Thanks also to Professors Robert Bracewell and Panayiota Kendeou for serving on my dissertation committee. Your comments and advice were much appreciated and contributed to the quality of the final product. Many thanks to Dr. Emmanuel Blanchard for translating the abstract into French.

Thanks go to the members of our research group, some whom have since moved on, for their ideas, encouragement, and general good-nature. A particular thanks to Carlos Nakamura for many long and stimulating conversations about research and your sense of humour.

I would like to thank my parents, who taught me to be independent and to take pride in my work. They have made this possible by always encouraging me to pursue my own goals and supporting my decisions. They have also been most generous, assisting me with the financial costs associated with my studies. Thanks to my family and close friends for understanding that I would have liked to have seen them more often than I was able. And a special thanks to Sonia Faremo for her patience and support during this unnecessarily long journey.

I would also like to acknowledge that financial support for this research was provided in part by the Herschel and Christine Victor Fellowship, a grant from the Social Sciences and Humanities Research Council of Canada INE Fund, and the James McGill Research Fund.
ABSTRACT

Authors augment their texts using devices such as bold and italic typeface to signal important information to the reader. These typographical text signals are an example of a signal designed to have some affect on others. However, some signals emerge through the unplanned, indirect, and collective efforts of a group of individuals. Paths emerge in parks without having been designed by anyone. Objects accumulate wear patterns that signal how others have interacted with the object. Books open to important, well studied pages because the spine has worn, for example (Hill, Hollan, Wroblewski, & McCandless, 1992). Digital text and the large-scale collaboration made possible through the internet provide an opportunity to examine how unplanned, collaborative text signals could emerge. A software application was designed, called CoREAD, that enables readers to highlight sections of the text they deem important. In addition, CoREAD adds text signals to the text using font colour, based on the group’s collective history and an aggregation function based on self-organising systems. The readers are potentially influenced by the text signals presented by CoREAD but also help to modify these same signals. Importantly, readers only interact with each other indirectly through the text. The design of CoREAD was greatly inspired by the previous work on history-enriched digital objects (Hill & Hollan, 1993) and at a more general level it can be viewed as an example of distributed cognition (Hollan, Hutchins, & Kirsh, 2000).

Forty undergraduate students read two texts on topics from psychology using CoREAD. Students were asked to read each text in order to write a summary of it. After each new student read the text, the text signals were changed to reflect the current group of students. As such, each student read the text with different text signals presented.

The data were analysed for each text to determine if the text signals that emerged were stable and valid representations of the semantic content of the text. As well, the students’ summaries were analysed to determine if students who read the text after the text signals had stabilised produced better summaries. Three methods demonstrated that CoREAD was capable of generating stable typographical text signals. The high importance text signals also appeared to capture the semantic content of the texts. For both texts, a summary made of the high signals performed as well as a benchmark summary. The results did not suggest that the stable text signals assisted readers to
produce better summaries, however. Readers may not respond to these collaborative text signals as they would to authorial text signals, which previous research has shown to be beneficial (Lorch, 1989). The CoREAD project has demonstrated that readers can produce stable and valid text signals through an unplanned, self-organising process.
RESUME

Les auteurs améliorent leurs textes en utilisant des signaux textuels tels que la mise en gras ou en italique pour informer le lecteur de la présence d’information importante. Ces marqueurs typographiques sont des exemples de signaux scientifiquement créés pour provoquer une réaction chez d’autres personnes. Cependant, il arrive parfois que des signaux émergent des efforts collectifs, indirects et non planifiés, d’un groupe d’individus. Des chemins peuvent émerger dans des parcs sans avoir été conçus par qui que ce soit. Les objets accumulent des traces d’usure qui illustrent comment leurs utilisateurs ont pu interagir avec eux. Les livres s’ouvrent parfois sur des pages importantes, fréquemment étudiées, car la reliure s’est usée (Hill, Hollan, Wroblewski, & McCandless, 1992). Les textes numériques et la collaboration à grande échelle rendue possible par l’Internet offrent une opportunité d’étudier comment des signaux textuels collaboratifs et non planifiés peuvent apparaître. Un logiciel, nommé CoRead, a été créé et permet aux lecteurs de surligner des sections jugées importantes d’un texte. De plus, CoREAD ajoute des signaux aux textes en colorant les fontes de caractères en fonction de l’histoire collective du groupe. CoREAD dispose également d’une fonction d’agrégation basée sur des mécanismes d’auto-organisation. Les lecteurs sont potentiellement influencés par les signaux transmis par CoREAD mais, dans un même temps, ils participent aussi à la modification de ces signaux. Il est également important de noter que les interactions entre lecteurs ne sont possibles que de manière indirecte, via le texte. La conception de CoREAD s’est fortement inspirée de travaux antécédents dans le domaine des objets numériques enrichis historiquement (Hill & Hollan, 1993). De manière plus générale, cette conception peut être vue comme un exemple de cognition distribuée (Hollan, Hutchins, & Kirsh, 2000).

En utilisant CoREAD, quarante étudiants sous-gradués ont du lire deux textes portant sur des sujets de psychologie. Il leur a été demandé de lire chaque texte de manière à pouvoir en écrire un résumé. Lorsqu’un nouvel étudiant lisait l’un des textes, les signaux de ce même texte étaient changés pour refléter le groupe actuel d’étudiants. Ainsi, chaque étudiant lisait un même texte qui comportait des signaux différents.

Les données furent analysées pour chacun des deux textes afin de déterminer si les signaux textuels qui émergeaient étaient des représentations stables et valides du contenu
sémantique du texte. De la même manière, les résumés des étudiants furent évalués pour
déterminer si la stabilisation des signaux textuels permettait aux étudiants de fournir de
meilleurs résumés des textes lus. Trois méthodes ont démontré que CoREAD était
capable de générer des signaux typographiques stables de texte. Les signaux textuels de
haute importance ont également semblé capturer le contenu sémantique des textes. Pour
les deux textes, un résumé obtenu en se basant sur ces signaux forts donnait des résultats
similaires qu’un résumé type. Cependant, les résultats n’ont pas montré que des signaux
textuels stables permettaient aux lecteurs de produire de meilleurs résumés. Il se peut que
les lecteurs ne répondent pas à des signaux textuels collaboratifs comme ils le feraient
pour des signaux textuels autoritaires dont l’intérêt a été démontré dans de précédentes
recherches (Lorch, 1989). Le projet CoREAD a mis en évidence que les lecteurs peuvent
produire des signaux textuels stables et valides dans le cadre d’un processus non planifié
d’auto-organisation.
# A self-organising systems approach to social software

## TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgements</td>
<td>i</td>
</tr>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Resume</td>
<td>iv</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>x</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xii</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td><strong>Chapter 1: Literature review</strong></td>
<td>3</td>
</tr>
<tr>
<td><strong>Complex Systems</strong></td>
<td>3</td>
</tr>
<tr>
<td>Simplicity, Organised Complexity, and Disorganised Complexity</td>
<td>4</td>
</tr>
<tr>
<td>Characteristics of Complex Systems</td>
<td>7</td>
</tr>
<tr>
<td>1. Many diverse agents</td>
<td>8</td>
</tr>
<tr>
<td>2. Interaction and stigmergy</td>
<td>8</td>
</tr>
<tr>
<td>3. Positive &amp; Negative Feedback</td>
<td>10</td>
</tr>
<tr>
<td>4. Emergence</td>
<td>11</td>
</tr>
<tr>
<td>5. Upward &amp; Downward Causation</td>
<td>12</td>
</tr>
<tr>
<td>6. History of the system</td>
<td>12</td>
</tr>
<tr>
<td>Summary of Characteristics of complex systems</td>
<td>13</td>
</tr>
<tr>
<td>Modelling complex social systems</td>
<td>14</td>
</tr>
<tr>
<td>Agent-based models</td>
<td>14</td>
</tr>
<tr>
<td>Schelling’s “racial segregation model”</td>
<td>15</td>
</tr>
<tr>
<td>The Standing Ovation Problem</td>
<td>16</td>
</tr>
<tr>
<td><strong>Empirical and computational studies</strong></td>
<td>17</td>
</tr>
<tr>
<td>The interaction of task difficulty and social networks</td>
<td>17</td>
</tr>
<tr>
<td>Confirmation bias at the group level</td>
<td>19</td>
</tr>
<tr>
<td>Diversity trumps ability</td>
<td>22</td>
</tr>
<tr>
<td>Summary of Empirical and computational studies</td>
<td>24</td>
</tr>
<tr>
<td>Summary of Complex Systems</td>
<td>25</td>
</tr>
<tr>
<td><strong>Social Software</strong></td>
<td>27</td>
</tr>
<tr>
<td>Some types of social software and their supported functions</td>
<td>27</td>
</tr>
<tr>
<td>Collaborative filtering and recommender systems</td>
<td>29</td>
</tr>
<tr>
<td>Collaborative bookmarking and tagging</td>
<td>30</td>
</tr>
<tr>
<td>Delicious.</td>
<td>31</td>
</tr>
<tr>
<td>CiteULike.</td>
<td>33</td>
</tr>
<tr>
<td>Summary of Social Software</td>
<td>36</td>
</tr>
<tr>
<td><strong>Social Software for Education</strong></td>
<td>36</td>
</tr>
<tr>
<td>CoFIND</td>
<td>38</td>
</tr>
<tr>
<td>Open University of the Netherlands Group</td>
<td>40</td>
</tr>
<tr>
<td>Altered Vista</td>
<td>42</td>
</tr>
<tr>
<td>AnnotatEd</td>
<td>43</td>
</tr>
<tr>
<td>OATS (Open Annotation and Tagging System)</td>
<td>44</td>
</tr>
</tbody>
</table>
Chapter 5: Results

Basic results characterising the study

Participants’ Prior Knowledge
  Flynn Effect text
  Problem Solving text
  Summary of Participants’ Prior Knowledge results

General outcomes
  Did participants use the highlighting function?
  Did the participants’ use of highlighting grow over time?
  Did the amount of text signalling increase over time?
  Were participants overly influenced by the collaborative text signals?
  Correlational analyses of basic measures
  Summary of general outcomes

Research Question 1 – Emergence of stable typographical text signals
  Percentage of the words that changed signal state (Method 1)
  Percentage of the words in the same signal state over eight participants (Method 2)
  Changes in the agreement between points in the history (Method 3)
  Results Summary: Research Question 1

Latent Semantic Analysis
  Creating a latent semantic “space”
  Using LSA to compute similarity
  Scoring essays with LSA
  Why use LSA?
  Were the LSA spaces used adequate?
  Flynn Effect text
  Problem Solving text
  A note about the Latent Semantic Analyses

Research Question 2 – Validity of the high signals
  LSA: high signals to the original text (RQ2.1)
  Flynn Effect text
  Problem Solving text
  Results Summary: Research Question 2.1
  LSA: high signals compared to a model summary (RQ2.2)
  Flynn Effect text
  Author’s overview as a model summary.
  Problem Solving text
  Author’s typographical text signals as a model summary.
  Author’s typographical signals compared to the collaborative typographical signals.
  Results Summary: Research Question 2.2
  Results Summary: Research Question 2

Research Question 3 – Better summaries post stability?
  Are post-stability summaries more semantically similar to the original text? (RQ3.1)
  How do the post-stability summaries compare to a model summary? (RQ3.2)
LIST OF TABLES

Table 1: Problems of Simplicity and Complexity* 5
Table 2: Mean (SD) performance (over 50 trials) of groups* 24
Table 3: Examples of social software 29
Table 4: Examples of social software for education 38
Table 5: Characteristics of text signals* 48
Table 6: Characteristics of the participants, the text, and the reading and recall conditions (Lorch et al., 1995) 49
Table 7: E-book reading devices* 52
Table 8: Comparing distributed cognition and complex systems 67
Table 9: Effects of the weighting function on the importance scores (text signal state indicated by colour) when only the first of five readers highlights a word 79
Table 10: Effects of the weighting function on the importance scores (text signal state indicated by colour) when only the first two of five readers highlights a word 79
Table 11: Text from the participant recruitment sign 86
Table 12: Participants’ demographic data* 87
Table 13: Data summarising the characteristics of both texts 89
Table 14: Verbal instructions: Reading Task 91
Table 15: Reading Task Instruction card 92
Table 16: Verbal instructions: Writing Task 92
Table 17: Writing Task instruction card 92
Table 18: Summary statistics (M and SD) for both texts [n=40] 94
Table 19: Mean (SD) percentage of the text displayed to the participants by signal state 100
Table 20: Mean (SD) percentage of text highlighted by signal state 102
Table 21: Statistically significant* correlations between basic measures for the Flynn Effect text 105
Table 22: Statistically significant* correlations between basic measures for the Problem Solving text 105
Table 23: Three measures of association (Percent agreement, Cramer’s V, Pearson’s correlation) for equidistant points in the history (by table row) – Flynn Effect text 111
Table 24: Three measures of association (Percent agreement, Cramer’s V, Pearson’s correlation) for equidistant points in the history (by table row) – Problem Solving text 111
Table 25: A ‘Near Neighbours Analysis*’: Top 10 terms most similar to the Flynn Effect text 116
A self-organising systems approach to social software

Table 26: A ‘Near Neighbours Analysis’: Top 10 terms most similar to the Problem Solving text 117
Table 27: Semantic similarity between the Flynn Effect text and the author’s overview, and two artificial summaries 119
Table 28: Semantic similarity between the Problem Solving text and the author’s signals, and two artificial summaries 120
Table 29: Similarity of the participants’ summaries to the texts by quartiles (n=10); measured by LSA cosine, M and SD reported. 124
Table 30: Comparison between the post-stability participants’ summaries and the model summary, for each text (LSA cosine measure) 124
LIST OF FIGURES

Figure 1: Tag cloud of popular tags from Delicious (Aug 30, 2008) 31
Figure 2: Bookmarking applications (see red box) displayed on a journal article website 32
Figure 3: The author’s personal CiteULike tag cloud (September 1, 2008) 33
Figure 4: Screenshot from CiteULike displaying an article in the collective library (August 31, 2008) 35
Figure 5: Screenshot of CoREAD 73
Figure 6: Course/Major topics studied related to the Flynn Effect text 95
Figure 7: Prior knowledge related to the Flynn Effect text 96
Figure 8: Course/Major topics studied related to the Problem Solving text 97
Figure 9: Prior knowledge related to the Problem Solving text 98
Figure 10: Percentage of the text in each signal state, as viewed by the participant, and the amount of text highlighted by the participant 101
Figure 11: Total number of words highlighted by each participant coded by signal state 103
Figure 12: Percent change in the collaborative text signals after each participant 108
Figure 13: Percentage of the collaborative text signals that remained the same over the last eight participants by state (low, moderate, high) and those that changed (unstable) 109
Figure 14: Binomial probability distribution for 45 trials when the probability of a match at each trial is 0.105 122
Figure 15: Semantic similarity between 1) the participants’ summaries, and 2) the model summary (indicated with a horizontal line) and the relevant text 126
INTRODUCTION

…the real problem-solving engine was the larger, biotechnological matrix comprising the brain, the stacked papers, the previous marginalia, the electronic files, the operations of search provided by the Mac software, and so on, and so on….ours are essentially the brains of natural-born cyborgs…

Andy Clark (2003, p.26)

This work began to take shape after a chance purchase of a book. The book, *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*, by Andy Clark (2003) advanced the extended mind hypothesis (Clark & Chalmers, 1998) by claiming that humans are “natural-born cyborgs”. By this, Clark suggests that the mind is naturally suited to augment itself through the use of external tools that store and process information. This is a position consistent with various perspectives of distributed cognition (Hollan et al., 2000).

The sixth chapter of the book, “Global Swarming”, introduced the idea of technologically-mediated mass collaboration among people, and was particularly responsible for the direction of this work. This swarm intelligence (Bonabeau, Dorigo, & Theraulaz, 1999; Bonabeau & Theraulaz, 2000) is a type of complex, or self-organising, system linking people through digital artifacts. The artifacts become augmented with socially derived metadata signalling the artifact’s usefulness, categories of interest, links to other artifacts, etc.

A study was designed to investigate the possibility that large groups of students could generate useful typographical text signals (e.g., using *italics* to emphasise words) through indirect collaboration using a self-organised process. Since self-organising systems are often adaptable, flexible, and robust the resulting text signals should be as well. These features could be useful in educational contexts where static, inflexible texts are more often the norm. Additionally, since the text signals are produced as a by-product of individual work the efficiency of this approach is very high. The text signals that emerge are free of any costs except the time it takes for a sufficient group of students to read the text. As such, instructors could add new texts in a timely fashion. This would be less likely if a long process involving expert preparation of the texts was required.
A review of the literature on complex systems, social software, text signalling and distributed cognition follows in order to situate this work within a broader research framework.

\[1\] Where the boundaries lie between distributed cognition and other related perspectives is difficult to say and so some scholars have begun combining these perspectives under the DEEDS acronym: dynamical, embodied, extended, distributed, and situated (Walmsley, 2008).
CHAPTER 1: LITERATURE REVIEW

Complex Systems

Maurice Maerterlinck, the Belgian poet, once wrote, “What is it that governs here? What is it that issues orders, foresees the future, elaborates plans and preserves equilibrium?” These, indeed, are puzzling questions.

Eric Bonabeau and Guy Theraulaz (2000, p.73)

A classic example of a complex system is a social insect colony, ant colonies for example (Bonabeau et al., 1999). In such collectives, rather simple individuals are able to solve difficult problems collectively (such as food foraging) in a decentralised fashion, that is, without a leader or a plan. Trails to nearby, rich food sources emerge from initially random individual searches of the environment. These random searches become organised as the ants respond to pheromones they leave in the environment, which creates a kind of positive feedback loop. A dominant pheromone trail emerges as more and more ants find the food source and return to the nest. This chemical trail is similar to trails made by people through parks since in both cases future activity becomes organised by past activity and the resulting trail was not designed, but emerged through decentralised activity.

Many other examples of complex systems could be given: slugs and their mucous trails (Clark, 2003), termite colonies (Bonabeau & Theraulaz, 2000), bee colonies (Seeley, 1995), neuronal activity and the central nervous system (J. H. Holland, 1995a, 1995b), and even cities (Jacobs, 1992).

Given that scholars who study complex systems come from a wide variety of disciplines and that the field is fairly young there are numerous terms used to refer to complex systems (also see Gell-Mann, 1995). These terms include self-organising systems (Bonabeau et al., 1999; Bonabeau & Theraulaz, 2000), complex adaptive systems (CAS) (J. H. Holland, 1995b; Miller & Page, 2007), dynamic systems (Howe & Lewis, 2005), and decentralised systems (Resnick, 1994). Complex systems with individually

---

2 There are many accessible books about complex systems available. A good, though basic introduction is Emergence: The connected lives of ants, brains, cities and software by Johnson (2001).
“intelligent”\textsuperscript{3} agents are sometimes referred to as swarm intelligence, social cognition\textsuperscript{4}, collective intelligence, etc.

Although the above example of an ant colony is helpful, we will require 1) a description of the boundaries that separate simple from complex systems, 2) the characteristics of complex systems, and 3) some ways that such systems are modelled. These are described below.

\textit{Simplicity, Organised Complexity, and Disorganised Complexity}

The importance of this middle region, moreover, does not depend primarily on the fact that the number of variables involved is moderate — large compared to two, but small compared to the number of atoms in a pinch of salt. The problems in this middle region, in fact, will often involve a considerable number of variables. The really important characteristic of the problems of this middle region, which science has as yet little explored or conquered, lies in the fact that these problems, as contrasted with the disorganized situations with which statistics can cope, show the essential feature of organization. In fact, one can refer to this group of problems as those of organized complexity.

Warren Weaver (1948), emphasis added

In 1948, Warren Weaver published an article that clearly described the three types of problems faced by scientists. He named these problems of simplicity, disorganised complexity, and organised complexity. These three problems and their characteristics are presented in Table 1. They are presented in the table with organised complexity in the centre to emphasise that it is the “middle region” that Weaver refers to in the quote above. However, they will be discussed below according to their historical development, which is how Weaver introduced them as well.

\textsuperscript{3} Where by the term “intelligent”, it is meant that the agents are capable of responding to their environment in some adaptive way. The agents could be people, insects, social animals, or even agents with artificial intelligence (e.g., collections of robots).

\textsuperscript{4} In this case, “social cognition” refers to a group acting in concert not an individual’s understanding of social interactions and rules, which is the usual usage in social psychology.
### Table 1: Problems of Simplicity and Complexity

<table>
<thead>
<tr>
<th>Properties</th>
<th>Simplicity</th>
<th>Organised complexity</th>
<th>Disorganised complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Few variables</td>
<td>• Many variables (often agents or entities interacting)</td>
<td>• Very many variables (or entities interacting)</td>
</tr>
<tr>
<td></td>
<td>• Can model with simple equations</td>
<td>• Some basic model of the agents is needed</td>
<td>• Little is known or needs to be known about the individual entities</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Ordered (hierarchic**)</td>
<td>• No order</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Higher levels are not merely the average of the lower levels (randomness does not “cancel out”)</td>
<td>• Randomness “cancels out”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Higher levels cannot be easily predicted from knowledge of the lower levels</td>
<td>• Can model aggregate or macroscopic behaviour through “averaging” lower levels (simple equations + probabilistic error)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Aggregate behaviour can be “predicted” only by running simulations of the lower level entities interacting</td>
<td></td>
</tr>
<tr>
<td>Examples</td>
<td>Classical physics $F = ma$</td>
<td>Ant colony</td>
<td>Molecular model of temperature: the average motion (energy) of all particles in the system</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Neurons in the brain</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Connectionist networks</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Information flow through social networks</td>
<td></td>
</tr>
</tbody>
</table>

* The three types of science problems were suggested by Weaver (1948), however, some of the table contents reflect the contributions of other complex systems scholars as well.

** Simon (1962; 1996) described complex systems as hierarchic, meaning that the interacting elements were “nearly decomposable” and organised or linked together. He did not mean to suggest a top-down traditional hierarchy with lower levels only connected to higher levels, such as a military chain of command.

Problems of simplicity, or simple systems, involve phenomena that have a few variables (or parts, agents, or entities). This characterises many of the problems addressed.
by the natural sciences, particularly the physical sciences, between the 17th and the 19th centuries (Weaver, 1948). Often these were two or three variable problems that could be modelled with simple equations such that if one variable were increased the other increased or decreased by some proportional amount. For example, force in classical physics is simply the product of an object’s mass and its acceleration. Increase either and force increases proportionally. Statistical analyses for such simple systems were only needed in so far as the measurements from actual experiments contained errors (Hacking, 1975)\(^5\). Progress on such problems was in fact made without much need of statistical theory or techniques. Solving the equations that resulted could be done analytically with algebra, or for variables changing in a regular manner over time with calculus.

Unlike problems of simple systems, problems of disorganised complexity did require statistical theory. In these systems there were very many variables or entities, often an essentially infinite number. For example, the number of oxygen molecules that would fit in an everyday container is a very large number. Rather than treat the molecules individually, an aggregate approach was taken whereby the whole could be treated as an average of the parts. The temperature of the air as a whole is treated as the average energy of the individual molecules.

\[\text{...in physics we can solve mathematically the two- and [infinite]- body problem, but no clean solutions exists for the intermediate cases. It is as if much of modern science counts using only 1, 2, and [infinity].}\]

---

John H. Miller and Scott E. Page (2007, p. 228)

Problems of organised complexity are those where there are many variables or entities, though perhaps not as many as in disorganised complex systems. The key difference, however, is that the entities are organised in some way. Because of the organisation they cannot be aggregated in a way that assumes independence – their actions cannot be averaged. With respect to the social sciences this is the “social science in between” (Miller & Page, 2007, Chapter 12).

Although Weaver (1948) used the term organised complexity to clarify his ideas, the simpler term complex systems is typically used and will be used in what follows. It

---

\(^5\) Unlike the social sciences “measurement error” in the physical sciences was considered strictly errors from the process of measuring phenomena, and did not include the idea of “unexplained” variance. That is, the models were considered correct and improvements in the measuring equipment itself were expected to result in smaller and smaller error variation over time. This indeed was often the case.
should also be pointed out that complex systems are not just complicated systems (Miller & Page, 2007, p. 8), like a car for example. Whereas both may have many parts that form a whole, in complicated systems the parts are basically independent. Remove a part and the effect it will have on the whole is fairly predictable (e.g., removing a car battery). In complex systems, the parts are interdependent and the effect of removing, adding, or changing a part does not allow one to make any simple predictions about the whole. In some cases the whole might be robust to the change and remain roughly the same, but in others the whole may be destroyed or completely changed. Some important characteristics of complex systems are described next.

**Characteristics of Complex Systems**

There are many good references that describe the characteristics of complex systems (Bonabeau et al., 1999; Bonabeau & Theraulaz, 2000; Elliot & Kiel, 2004; Heylighen, 2001; J. H. Holland, 1995a, 1995b; Howe & Lewis, 2005; Kennedy, Eberhart, & Shi, 2001; Marsh & Onof, 2008; Miller & Page, 2007; Mitchell, 2006; Resnick, 1994; Sulis, 1997; Weaver, 1948). Although there are differences among the perspectives regarding what constitutes a complex system, the six characteristics discussed below are fairly agreed upon for complex systems where the agents are capable of responding to their environment in an “intelligent” manner. These complex systems of intelligent agents can be contrasted with some complex systems of physical phenomena where the agents are molecules (or something similar) and have a very limited ability to respond to their environment. For example, a material like iron that can be magnetised can be considered a collection of small magnets, or spins (see Heylighen, 2001). As the temperature decreases these spins tend to align such that the whole piece of iron becomes magnetised; increasing the temperature causes the spins to point in random directions and the iron ceases to be a magnet. Although the spins self-organise and become ordered they have a limited ability to respond to the environment. They also lack diversity, except in that they initially point in different directions.

---

6 Some scholars prefer the term “complex adaptive systems” for complex systems that have the potential to adapt because the agents themselves are more sophisticated (i.e., are capable of learning) and can also interact (i.e., communicate) in more sophisticated ways (J. H. Holland, 1995a, 1995b; Miller & Page, 2007).
The six characteristics of complex systems described below apply to systems with intelligent and diverse agents, such as social insects and animals, and of course people.

1. Many diverse agents

Perhaps the most important characteristic is that the system is composed of many agents. As well, these agents are diverse either with respect to their own skills and knowledge (e.g., people) or their current direction of travel, location in the environment or task (e.g., ants). Some of this diversity is essentially due to randomness or noise. For example, the direction in which ants initially set out to forage for food is random and only later becomes organised as will be discussed below. Additionally, the ants do not collaborate perfectly; they follow pheromone trails imperfectly. This error or “noise” allows for constant exploration (Resnick, 1994, p. 136-9). For people, diversity comes from each individual’s knowledge and skills. This diversity can be considered as random, in that a group of agents has been randomly sampled from a larger population.

The presence of a large collective of diverse agents permits the group to explore the problem space more thoroughly at the collective level than is possible at the individual level. Each individual then can act based on his exploration of the problem space or by exploiting the solutions of others. Exploiting the solutions of others requires that the agents share solutions (partial or complete) with others through interaction.

2. Interaction and stigmergy

To benefit from the group’s collective intelligence the agents must interact, either directly or indirectly. Direct interaction occurs when two agents share information with one another. For example, two colleagues sharing notes or ideas. Indirect interaction can occur in a few ways. For example, a network7 defining the direct connections between agents also permits information to travel indirectly. Agent A communicates directly with agent B who in turn relays the information to agent C; communication between agents A and C is indirect, for example. Another form of indirect collaboration occurs through what is known as stigmergy (Bonabeau et al., 1999; Dron, Boyne, & Mitchell, 2001; O. Holland & Melhuish, 1999; Marsh & Onof, 2008; Sulis, 1997; Theraulaz & Bonabeau,

---

7 There are many standard types of networks that are used to model such interactions (e.g., lattice, small word, fully connected, random, etc.) (Goldstone, Roberts, Mason, & Gureckis, 2008; Miller & Page, 2007; Mitchell, 2006)
Stigmergy originally referred to products of work left in the environment that influenced future work (Grasse, 1959). An example is the building behaviour of termites whereby termites are influenced by current structures to add to those structures. The result is the creation and maintenance of the termite colony’s nest.

La coordination des tâches, la régulation des constructions ne dépendent pas directement des ouvriers, mais des constructions elles-mêmes. L’ouvrier ne dirige pas son travail, il est guidé par lui. C’est à cette stimulation d’un type particulier que nous donnons le nom de STIGMERGIE (stigma, piqûre; ergon, travail, œuvre = œuvre stimulante).

Pierre-Paul Grassé (1959, p.65), emphasis in the original

The coordination of tasks and the regulation of constructions does not depend directly on the workers, but on the constructions themselves. The worker does not direct his work, but is guided by it. It is to this special form of stimulation that we give the name STIGMERGY (stigma, goad; ergon, work, product of labour = stimulating product of labour)

translation provided by Owen Holland and Chris Melhuish (1999, p.1)

The term stigmergy has also been extended to include signs left in the environment that are not considered products of work. The pheromones that ants deposit in the environment are a classic example. These chemicals help organise the ants behaviour but are not themselves part of the task; foraging for food, for example.

The key to any stigmergic sign is that it is present in the local environment (physical or symbolic) in which the agents “work”. Agents only have access to this local information and do not have access to the entirety of the group’s efforts. Also, the signs in the environment are readily perceived and interpreted by the agents.

Human signs, in symbolic environments like written documents are also in some sense “physical”. Therefore these signs can be;

1. a “physical” modification of the text environment (e.g., a note or a rewrite),
2. a sign that does not necessarily add or change the text (e.g., a highlight) but signals something about the text, or
3. both (e.g., combinations of marks that indicate importance or categorise information, and also add content in the form of notes).

---

8 Marsh and Onof (2008) have also suggested the term stigmergic cognition to refer to indirect communication mediated by modifications to the environment – a form of extended mind (Clark, 2001; Clark & Chalmers, 1998).
It should be noted that in the case of stigmergic, and therefore indirect, collaboration no specific network of links between the agents need be specified. Agents do not interact with some specific subset of the group but rather interact with the signs left behind by some unknown and constantly changing subset of the group. At any given point there may be an identifiable network, but this network is constantly changing moment to moment as signs are added to or fade from the environment.

The interaction among the agents means that the agents are not independent. To the extent that the future actions of some agents are dependent on the previous actions of other agents, positive feedback ensues.

3. Positive & Negative Feedback

Positive feedback is particularly important in complex systems. It can be described as when a given action increases the likelihood of the same action occurring again. This results from agents exploiting the knowledge of other agents. Positive feedback tends to move a system towards a new state by amplifying actions. As such, positive feedback is particularly important in complex systems because it allows the collective to pursue solutions that seem promising. However, if all agents simply began following a single solution the collective would only perform as well as that one potential solution. Therefore, a balance between the exploitation of promising solutions through positive feedback and individual exploration (which adds new information) or some form of negative feedback is needed.

Negative feedback can be described as actions that maintain the current state or return the system to a previous and desired state. A thermostat is a classic example of negative feedback. As the temperature moves away from a desired level the thermostat acts to return the temperature to the desired state by turning on a heating or cooling device. In a complex system such as an ant colony the rate at which pheromones in the environment fade away can be considered a kind of negative feedback. As the pheromones fade the stigmergic signs in the environment return to zero unless new pheromone deposits are made. This permits pheromone trails to diminishing food sources to weaken and encourages members of the colony to seek out food elsewhere.
If interaction through positive feedback, tempered by individual exploration and negative feedback, yields new and stable outcomes then outcomes emerge at the collective level from the interaction of the lower-level agents.

4. Emergence

Emergence is a particularly difficult concept to define, or even recognise in particular situations. An emergent phenomenon is one that is a stable pattern, occurs at a level above the agents themselves (at the collective level) and should not be directly predictable from agent level behaviour (not just more of or an average of lower level behaviour). Often, it is adaptive at the collective level. The difficulty in assessing whether emergence has occurred arises from the difficulty in assessing stability, particularly since complex systems are generally constantly undergoing change and may stabilise more than once at different times and different configurations.

An example from evolution may clarify this. Basic evolutionary theory posits that species change over time in response to natural selection. That is, species change in ways that are better adapted to the environment. A question arises though about when speciation actually occurs. How similar must a group of animals be to each other and how different must that group of animals be from another (ancestral) group to be considered a new species? Or when has a new species emerged from evolutionary processes?

In many cases, scholars have relied on both the human ability to perceive stable patterns in visual displays of the organised collective behaviour or outcome (Bonabeau & Theraulaz, 2000; O. Holland & Melhuish, 1999; Schelling, 2006) and simple metrics that measure relevant properties of the system state. For each approach, the system is “viewed” at different times in the system’s history to determine if, when, and how the system stabilised. For example, social insects like ants remove their colony’s dead and create cemeteries. Creating an array of photographs taken over time permits one to easily recognise when, where, and how the piles were created (Bonabeau & Theraulaz, 2000, p.78). A similar approach was used to present the results of a simulation of this sorting behaviour with robots (O. Holland & Melhuish, 1999).

Phenomena developing at higher levels (or stable emergent phenomena) also affect the lower-level agents. Of course, this developing collective phenomenon is a product of the interactions of the lower-level agents. The agents modify the environment, other
agents then modify the environment in part in reaction to the previous modifications, and so on. This begs the question “which level is causing the other?” The answer, though uncomfortable, is that there is both upward and downward causation.

5. Upward & Downward Causation

Since the agents and the environment – which stores their developing collective opinion – are in constant interaction there is no sense in which one simply causes the other. Rather there is an ongoing back and forth, upward (agent to collective) and downward (collective to agent) interaction. This is somewhat foreign to the traditional way of thinking about causes and effects with an emphasis on independent and dependent variables.

Given upward and downward causation, and how higher-level phenomena develop with periods of instability and stability, the history of the system as it evolves over time is critical.

6. History of the system

One thing that should be clear from the discussion of upward and downward causation is that the agents in complex systems are interdependent. The future state of the system will also be dependent on previous states, and these will be non-linear dependencies (Howe & Lewis, 2005). As such, there is no simple way to predict the future states from knowing previous states. In empirical research or simulations of complex systems, the system must be allowed to progress until a certain level of stability is reached or a certain period of time has elapsed. Predictions may be made based on observing several runs of the system (empirical) or model of the system (simulation). Statistical analyses of the sample of runs of the system may also prove useful in this regard.

Additionally, one cannot predict the emergent phenomenon from an understanding of the agents’ behavioural rules. The interdependencies among the agents produce collective phenomena that are more than a mere average or straightforward combination of the individual behaviours. For example, the collective behaviour of ants foraging for food is determined by the quality of food sources in the environment, the pheromone levels deposited by the ants, the rate at which these pheromones decay, the rules
specifying how the ants respond to the pheromones, etc. Changes to some of these conditions may have no effect on the colony’s general ability to locate the optimal food source. Some changes will result in the ants foraging almost entirely independently (e.g., very rapidly decaying pheromones that provide no stigmergic signs after a short time) while others will result in the colony exploiting sub-optimal food sources. How combinations of these conditions will affect the results is generally not known prior to empirical study or running simulations using models of the system.

Summary of Characteristics of complex systems

To recap, complex systems have several key characteristics. They have many diverse agents operating without a central authority. Higher level emergent phenomena are created through the system’s self-organisation driven primarily through positive feedback. This feedback entails both upward and downward causation making standard prediction techniques inapplicable. Because of these characteristics complex systems are often adaptive, flexible, and robust (Bonabeau et al., 1999):

1. Adaptable – solutions to problems emerge over time and therefore can adapt to changes in the context of the problem or the available solutions.
2. Flexible – the systems are “designed” to solve a class of problems (e.g., food foraging) and can respond to a variety of specific situations.
3. Robust – removing or changing some of the individual agents in the system generally does not compromise the performance of the collective.

Given the characteristics discussed, it is not possible to model complex systems using traditional methods such as using simple equations relating a few variables that can be solved analytically (Miller & Page, 2007). Even the addition of random error terms to such approaches does not solve this problem since the agents’ behaviours are not independent, hence neither are the errors (which do not simply “cancel out”). As such, a different modelling approach is needed. The most common approach for modelling social systems is agent-based modelling, discussed below.
Modelling complex social systems

[Agent-based models] overcome an assumption that underlies much of cognitive science – that the individual is the crucial unit of cognition. The alternative advocated here is that individuals participate in collective organizations that they might not understand or even perceive, and that these organizations affect and are affected by individual behavior.

Robert L. Goldstone and Marco A. Janssen (2005, p.424)

Complex systems can be modelled in a variety of related ways: connectionist networks, cellular automata, or agent-based models, for instance (Bonabeau, 2002; Elliot & Kiel, 2004; Goldstone & Janssen, 2005; Miller & Page, 2007; Page, 1999). There are three features shared by these models. First, there is some type of agent whether they are simply the nodes in a connectionist network, cells on a rectangular grid in cellular automata, or agents in an agent-based model. These agents follow a set of rules that specify their behaviour, often simply condition-action rules. Second, these agents are connected in some way to other agents. For example, the links connecting the nodes in a connectionist network, cells linked only to adjacent cells in cellular automata, or a social network that specifies how the agents are related in an agent-based model. In cases where agents interact only indirectly through stigmergic signs left in the environment, these connections between agents are replaced by “patches” in the environment that hold information, agents interact directly with these patches (Resnick, 1994). Third, the agents execute their actions over many iterations of the model (i.e., over time) either in parallel or in some sequence (in a specified or random order).

Agent-based models

Agent-based models have been the focus of most of the efforts to model complex social systems (Bonabeau, 2002; Elliot & Kiel, 2004; Gilbert, 2008; Goldstone & Janssen, 2005; Miller & Page, 2007; Page, 1999). These models are more suited to social interaction since they allow for agents with more complex rules governing their behaviours and various ways of linking the agents (i.e., a variety of social networks and stigmergic interactions). The benefits of using agent-based models include (also see Miller & Page, 2007, Chapter 6):

---

9 Conway’s ‘Game of Life’ is the most well known example of a cellular automata model (see Miller & Page, 2007, p. 52)
1. Control over the rules governing the agents’ behaviours.
2. The ability to manipulate the diversity of the agents by manipulating the proportion of agents following different rules.
3. The possibility of having agents change their behaviours through learning mechanisms during a model run. The agents themselves can adapt.
4. Control over the social network or spatial arrangement of the agents that defines the interactions they may have with other agents (direct interaction or indirect through the environment).
5. The possibility of changing task conditions during a run of the model to observe how the system adapts to the new problem.
6. The ability to rerun the model many times under the same or different conditions to observe general outcomes. This is also cost effective, unlike repeating empirical studies under the numerous conditions possible.
7. The ability to provide a precise description of the microconditions (agents, their interaction patterns, etc.) that lead to some macrophenomenon.

For rebuttals to common criticisms of computational models, in relation to agent-based models and social science, see Miller and Page (2007, Chapter 5). Some examples of agent-based models are presented below.

*Schelling’s “racial segregation model”*

Perhaps one of the first examples of modelling a complex social system was Schelling’s “racial segregation model” (Schelling, 2006). The model examines how a community’s level of segregation might emerge from the choices of the individual people. It is extremely simple: there are two types of agents (X’s and O’s for instance), agents have simple behavioural rules, and they are located on a simple two dimensional grid representing each agent’s location in a community. Agents are linked to the eight agents surrounding them on the grid – this is their current neighbourhood. Each agent has a desired minimum level of sameness within his neighbourhood (i.e., the minimum percentage of X’s or O’s depending on his group membership). An agent will move to a

---

10 Generally, this means that the agents are only directly connected to some subset of the remaining agents, though complete connectivity is also possible (though it often results in poorer performance).
vacant location that satisfies this criterion if his current neighbourhood does not, otherwise the agent does not move. Such models are run until all agents are satisfied with their neighbours – the system is completely stable. The interesting finding is that the level of segregation that usually emerges is much higher than the minimum level that any of the agents desired. Low levels of desired sameness, not an unreasonable assumption about human behaviour, can lead to extreme levels of segregation. This is an example of how averaging the behaviour of the individuals does not lead to either a proper understanding of the process, or even a correct estimation of the outcome at the aggregate level.

The reader will note that although the system is complex because of the interdependencies among the agents in the model\textsuperscript{12}, the agents themselves are quite simple. There is a worry that some agent-based models, for modelling collective human behaviour, may be in error due to unfounded assumptions about the agents’ abilities. This has led to an interest in combining cognitive models (or architectures) with agent-based models, thereby giving agents more realistic individual behaviours and abilities (Sun, 2006).

\textit{The Standing Ovation Problem}

Agent-based models are not analytic models. There are \textit{no} simple equations\textsuperscript{13} that can be solved, rather the model needs to be “run” in a simulation. Miller and Page (2007) provide an excellent example of this difference with the Standing Ovation Problem. An analytic approach might assume each audience member stands if his enjoyment level reaches a certain threshold, these might be assigned to each member following a distribution of some sort (the normal distribution, for example). The standing ovation phenomenon would then be examined by merely generating the proportion of the audience that reached the standing threshold. There is no interaction among the members of the audience in this case.

\textsuperscript{11} Schelling (2006) was careful to note that racial segregation is often due to other factors such as centralised control and differences in wealth and economic opportunity.

\textsuperscript{12} When any agent moves it changes the neighborhoods of several other agents as well, both at the location it left and the new location to which it moved.

\textsuperscript{13} By simple, it is meant that analytical approaches cannot mathematically model situations where there are many diverse agents interacting in nonlinear ways. Rather, when such approaches are used for many agents, the agents are all the same (homogeneous). The agents also do not interact, or do so in simple linear ways. This allows for solutions based on the “average” behaviour of the group of agents.
Clearly, this is probably not how standing ovations occur. Rather, the members of the audience influence one another — they are not independent. The decision to stand, and when this occurs, is probably dependent on the person’s opinion (as before) as well as her location in the theatre, what her companions decide to do and when, etc. In this model, there is no analytic solution. The collective behaviour cannot be computed from variables describing individual preferences, even with the addition of an error term. Rather, the model needs to be “run” in a simulation with many agents who each have rules describing how to assess a performance and how influence from strangers and friends affects the agent. Also, information about where each agent is sitting (to determine which people can influence the agent) and the pre-existing relationships between the agents will be needed. The model, when run, will yield a time course of activity across the audience members. If a standing ovation forms, when it started, and where it started will all be revealed in the model run. Computing averages of the same model, run many times, will often yield useful information\textsuperscript{14}. Namely, an audience with a certain composition, of a given size, etc., yields standing ovations a certain percentage of the time when the performance reaches a certain level.

\textit{Empirical and computational studies}

\textit{The interaction of task difficulty and social networks}

Goldstone and colleagues (Goldstone & Ashpople, 2004; Goldstone et al., 2008; Gureckis & Goldstone, 2006) have conducted some empirical studies to examine how fairly large groups of people (e.g., 15 individuals) can solve problems collectively. In some cases they have also developed models of these studies using agent-based models. These studies have examined path formation (Gureckis & Goldstone, 2006), foraging for resources in a virtual environment (Goldstone & Ashpople, 2004), and search for optima in an abstract search space\textsuperscript{15} (Goldstone et al., 2008; Gureckis & Goldstone, 2006).

In one study, individual performance on a search task was examined under four different network conditions and two levels of difficulty (Goldstone et al., 2008). The task involved picking a number from 0-100 that would return a payoff based on a payoff

\textsuperscript{14} Note that the sample in this case is the runs of the model (repetitions of the aggregate behaviour or macrophenomenon) not the individual audience members involved.
function. The easier task used a unimodal payoff function having one global optimum (and no local optima). The harder task used a trimodal payoff function, one global optimum, and two local optima. The participants received feedback about their choices on each round (the payoff plus some error) over 15 rounds. Additionally, they also received information about the numbers selected and associated payoffs for other participants in their network. The four network structures used were lattice, small-world, random, and fully connected. The lattice network is a circle where each individual is connected to two other people. The small-world network is the same except that there are a few additional links added. These links provide shortcuts through the network so that some individuals are connected to more than two people, one of whom is far from their location on the circle. The random network has random links, with the additional constraint that there is some path from every individual to every other one. In the fully connected network every person is connected directly to everyone else. Therefore, in this network each individual received information about how every other person in the group performed on every round.

Twelve groups of undergraduate students (median group size was 14) completed all eight combinations of difficulty and network (i.e., it was a fully within design). The dependent variable was the percentage of members of the group that were within one standard deviation of the global optimum.

For the easy task (unimodal function), the fully connected network results in better performance in the early rounds – more individuals approach the global optimum more quickly. However, later on the small-world and random networks also perform as well. For easy tasks, more information from one’s peers leads to faster improvement in individual performance. This is because there are no local optima that the group can converge on through sharing information. Moving into a local optimum early would cause the group to have trouble moving out of it later since few if any members would be searching other areas of the problem space.

For the difficult task (trimodal function) the small world function performed better in the early rounds (1-6). The fully connected groups quickly converged on local optima and got stuck there for a time. More information is not better. The small-world network

---

15 Also called “propagation of innovations” (Gureckis & Goldstone, 2006).
performs better because there is a balance between the exploration of the space by the members of the group and exploiting the better solutions through the sharing of information. It should be noted that the small-world and fully connected networks produced equivalent results in later rounds. Eventually, even the lower levels of exploration supported by the fully connected network moved the group out of the local optima and most members located the global optimum.

These results indicate that harder problems are more easily solved when groups more fully explore the problem space. This may result when a group consists of diverse members, each of whom uses different search strategies or who represent the problem in different ways. Exploration may also be promoted by the type of network that links the members of the group. More generally, this means that limiting communication, which also limits an individual’s ability to exploit (copy) the best solution found so far, promotes individual exploration of the space. Sharing some results after individual exploration takes place can often be better for hard problems than full sharing early on. Goldstone and his colleagues (Goldstone et al., 2008) also created a model of this phenomenon. Agents picked a number either based on their previous guess, by using their neighbours’ best guess, or by exploring the problem space. The probability of choosing a strategy was based on the agent’s pre-existing bias (modelled quantitatively) and the observed successfulness of the strategy. This model was run using groups of 15 agents guessing over 15 rounds. The model was run 1000 times and the results averaged. The pattern of results produced by this relatively simple model replicated the empirical results.

Confirmation bias at the group level

Hutchins (1995a) has also investigated the effects of group communication on group problem-solving using a computational approach. In this case, the group of agents was quite small (six agents). Each agent was modelled using a constraint-satisfaction network and the group was modelled by linking these networks into a “community of networks”.

---

16 The neighbours were the other agents that were directly connected to the agent in the network.
In a constraint-satisfaction network, there are several hypotheses (the nodes) that are linked in a connectionist network. Some hypotheses support each other and so their links are reinforcing (positive) while others are inhibiting (negative). The hypotheses also have initial activation levels simulating prior knowledge about the world\textsuperscript{17}. The hypotheses also receive input activation based on information from the world. Running the network will eventually lead to it settling on some pattern of activation representing the hypotheses the agent considers true and those that are false about the world. This is the network’s, the agent’s, interpretation of the situation.

Hutchins notes constraint-satisfaction networks do not easily change activation levels after once settling on a stable pattern of activation. That is, once this settling has occurred the network is less able to change activation patterns given new information from the world. As such, these networks are capable of modelling confirmation bias. Although a single individual modelled using a constraint-satisfaction network may display confirmation bias would a group of individuals collaborating also be prone to this bias?

\begin{quote}
\text{…one wonders if it might not be possible for a group of individuals, each of whom has this propensity [for confirmation bias], to make a different sort of trade-off. That is, might a group be organized in such a way that it is more likely than any individual alone to arrive at the best of several possible interpretations.}
\end{quote}

\begin{quote}
\text{…some ways of organizing people around thinking tasks will lead to an exacerbation of [confirmation bias], whereas other forms of organization will actually make an adaptive virtue on the group level of what appears to be an individual vice.}
\end{quote}

Edwin Hutchins (1995a, p.240)

To form a group, or a community of networks, Hutchins connected the individuals (the networks) to each other at the level of the hypotheses in their individual networks. This means that the individuals were sharing ideas about each hypothesis as they were trying to interpret the situation. This is in contrast to each network settling on an interpretation and then sharing the whole “answer”. The extent to which one individual influenced another individual was manipulated by the weight assigned to the link between their hypotheses. This models the persuasiveness of the communication between the individuals. A weight of 0 means no communication took place, or rather it was

\textsuperscript{17} Although not expressed by Hutchins (1995) this initial activation might well be considered the initial plausibility of the hypotheses for the individual.
completely unpersuasive. In the simulated groups, all six individuals were completely connected to everyone else. Each began the simulation with a different pattern of activation across the hypotheses. All of the individuals began the simulation undecided between the two possible interpretations, approximately at the halfway point between them. The hypotheses in the constraint-satisfaction networks and their connections were based on an actual case study. This case involved a collision between two ships. The group (community of networks) created models the communication between a captain and his crew members. In the actual case, this crew caused the accident to occur because the captain misinterpreted the situation and received no corrective information from his crew (who believed he held their correct interpretation of the situation). Although this work is based on a simplified model, it is a model of a real situation.

For a group of six individuals who did not communicate at all (persuasiveness set to 0 for all between-network links) three individuals settled on one interpretation and the other three on the other interpretation. With high persuasiveness all members of the group rapidly converged on one single interpretation. This outcome is similar to confirmation bias since the decision is made quickly and prevents new information from affecting the interpretations made. This, however, is a form of group-level confirmation bias. In other models where persuasiveness started out low then increased over time the individuals did not rapidly converge on a single interpretation. Rather, some individuals initially moved strongly to one interpretation but later the group reached a consensus at a single interpretation.

Hutchins argues that these findings indicate that groups can avoid group-level confirmation bias in cases where the individuals do not share everything they know continuously. That is, where high-bandwidth communication is not the norm. This of course is counterintuitive. Typically one thinks of groups performing better if they can share more information with one another. However, in the context of self-organising systems this result is not unexpected. A system of agents needs a balance between individual exploration of the problem space, which brings diversity to the possible set of solutions, and exploitation of promising solutions already found. High-bandwidth communication is likely to prematurely end exploration, as all agents tend to adopt the best current solution (since they all know about it). This can cause the group to converge
on local optima as a solution rather than continue searching for the global optimum. These findings, based on a computational model, are very similar to those found by Goldstone and colleagues discussed above (Goldstone et al., 2008; Gureckis & Goldstone, 2006).

Individuals prone to confirmation bias can, as a group, avoid this outcome when they share just enough information so that they individually continue to explore the problem space. For hard problems, the right type of communication can let a group elevate the performance of the individuals within it. The idea that groups can outperform individuals under certain circumstances has also been investigated by Page and Hong (Hong & Page, 2004; Page, 2007) using computational studies with agent-based models.

**Diversity trumps ability**

In working through the implications of my model, I stumbled on a counterintuitive finding: diverse groups of problem solvers – groups of people with diverse tools – consistently outperformed groups of the best and the brightest. … In my model, *diversity trumped ability.*

Scott E. Page (2007, p. ix-xx), emphasis in the original

Page (2007) argues that a randomly selected group of diverse and smart individuals perform better than a group of the highest-ability individuals\(^{18}\). With Lu Hong, he has produced a mathematical proof for this claim as well (Hong & Page, 2004). The claim rests on a few assumptions:

1. The problem is difficult. No single agent can solve the problem all of the time.
2. All of the agents have some ability to solve the problem; they can at least locate some of the local optima.
3. The population of agents to choose from is diverse enough that there always exists one who can improve on the current solution (until the global optimum has been reached)\(^{19}\).

---

\(^{18}\) High-ability individuals chosen from the higher end of a distribution are by definition less diverse. Though in more realistic situations, where individuals cannot be distributed on a single dimension, this may not exactly be true.

\(^{19}\) Note that Page acknowledges that this condition is probably too strong and indicates that in many of the simulations he ran it was not met but the ‘diversity trumps ability’ outcome still occurred (Page, 2007, p. 165). As such, the condition may be restated as “the population’s best solutions must not all contain low values”.
4. The set, or population, of agents from which the groups will be drawn is large and the groups created must not be too small. Given the above conditions Page states that “a randomly selected collection of problem solvers outperforms a collection of the best individual problem solvers” – this is the Diversity Trumps Ability Theorem (for the mathematical theorem see Hong & Page, 2004; for the quote see Page, 2007, p. 162).

The computational studies by Hong and Page used a very similar task to the one used by Goldstone described previously (Goldstone et al., 2008). Agents searched a space of 2000 points, represented as a circle. Each point had a value ranging from 0-100 (of real numbers not integers) and it was the agent’s task to locate the highest valued point. Each agent had a unique and very simple search algorithm. It searched three specific points to the right of its current location on the circle and retained the one with the highest value. These three points were limited to a distance of 12 beyond the current position. This results in 1320 unique agents (the population of agents).

Groups of agents worked together in sequence. The first agent would return the best point on the circle and the second agent would continue from there. This would continue until the group could not improve on the result: the group had become stuck in a local optimum (or had found the global optimum).

To compare diverse groups to groups of best individuals each agent was evaluated. The average performance of the agent across all starting positions on the circle was calculated, and the agents rank ordered. Groups of 10 and 20 agents were then created. Diverse groups were selected at random. Best agent groups were formed with the 10 best agents in the entire population. For each group size the average results for 50 trials were compared.

For both group sizes, the average performance (over 50 trials) of groups of agents picked at random was statistically significantly higher than the groups of best agents (see Table 2 below). Both types of groups performed well on average – achieving values of over 90 where the global optimum is 100. Although the difference between the types of groups appears to be small, the effect sizes are very large (see Table 2). Measures of the diversity of the groups were also computed and as would be expected the random groups were much more diverse than the best agent groups. The diversity of the groups of best
agents increased with group size (i.e., 10 to 20); there was less of an increase for the random groups.

**Table 2: Mean (SD) performance (over 50 trials) of groups**

<table>
<thead>
<tr>
<th>Group size</th>
<th>Type of Group</th>
<th>Best agents</th>
<th>Random agents</th>
<th>Effect Size ($d$) **</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Best agents</td>
<td>92.6 (0.02)</td>
<td>94.5 (0.01)</td>
<td>95</td>
</tr>
<tr>
<td>20</td>
<td>Random agents</td>
<td>93.8 (0.02)</td>
<td>94.7 (0.01)</td>
<td>45</td>
</tr>
</tbody>
</table>

*Results from a computational study by Hong and Page (2004)

**Cohen’s $d$ was calculated using the larger of the two standard deviations. The effect size was not reported by the study authors.

These results indicate that diversity is an important variable in the performance of a group of agents. First, randomly selected groups, which are more diverse, outperform groups of best agents. This is quite counterintuitive. Forming groups with agents that perform the best individually would lead one to expect the best performance possible at the group level. Second, increasing the group size tends to increase diversity, but this effect is more pronounced for the group of best agents. Because of this, the performance advantage that the randomly selected group has over the group of best agents decreases – the effect size is reduced by over half when there are 20 agents per group.

Some concerns about this computational study are that the computational model is based on a rather artificial task and very simple agents – the search algorithms they use are simple. As well, the agents do not have any knowledge about the problem space and therefore do not select strategies suited to it. Finally, the agents do not learn, or adapt their own strategies, from their search of the space. In any case, these results are interesting because they are counterintuitive.

**Summary of Empirical and computational studies**

In the cases examined, difficult tasks were more successfully completed when the members of the group balanced *exploration* and *exploitation*. That is, the members of the group did not copy the best solution early on but continued to add new solutions to the pool by “individually” exploring the problem space. Exploration is supported by two characteristics of groups. First, there must be some diversity within the group that originates from differences among the agents. These may be pre-existing differences in
knowledge, or differences originating from random exploration of the space. Second, communication among the agents must not be too rich early on in the problem solving process. These communication levels can be modelled by the density and strength of the links among the agents: the social network. Too much communication early on can often lead to the group finding and getting stuck in a local optimum; a kind of group-level confirmation bias.

*Summary of Complex Systems*

A collective intelligence consists of a large number of quasi-independent, stochastic agents, interacting locally both among themselves as well as with an active environment, in the absence of hierarchical organization, and yet which is capable of adaptive behavior.

William Sulis (1997, p. 35)

Up until this point, I have been using the term complex systems since it is widely used and understood. However, the software designed (see Chapter 2 below) was based on a self-organising system, modelled on the indirect interactions of ants through the pheromone trails left in the environment (Bonabeau, 2002; Bonabeau et al., 1999; Bonabeau & Theraulaz, 2000). This is a form of sign-based stigmergy. A short summary of the key features of self-organising systems – with respect to this study – is presented below:

1. There are many individual agents acting, often simultaneously\(^{20}\), using rather simple rules.
2. The agents act on the local information in their environment.
3. The environment plays a role by storing information.
4. The collection of agents is able to explore multiple solution paths, often through “random” exploration by individual agents.
5. Positive feedback amplifies actions such that more and more agents engage in the same action, negative feedback counterbalances this tendency.

Ultimately, global level phenomena emerge from these local interactions - the agents’ collective behaviour becomes organised although there was no guiding plan or leader coordinating their behaviour.

---

\(^{20}\) Simultaneous, or synchronous, interaction is the norm in natural systems like ant colonies. However, sequenced, or asynchronous, interaction is also possible.
These complex, self-organising systems can often be modelled using agent-based models as well as other modelling approaches. Empirical and computational studies have suggested that under certain circumstances groups outperform individuals. This tends to be the case for difficult problems when there is a balance between exploration of the problem space and exploitation of possible solutions by the group of agents. The performance that emerges from the interaction of the individuals is different and superior to the “average” of the individuals’ performances. The studies reviewed focused mainly on relatively simple problems or simplified models of real cases (Hutchins, 1995a). To what extent do groups, even very large distributed groups, outperform their best members in realistic everyday problems?

Self-organising systems typically require fairly large groups of diverse individuals. Bringing larger groups of people together efficiently has often been difficult given communication and organisation costs. In educational settings, for example, groups are often limited to three to five students. As another example, online discussion forums do not scale well as the number of participants and postings increases. Eventually too many posts and too many threads accumulate. Extracting anything useful from the forum becomes a challenge in itself. However, new forms of internet-based software permit vast numbers of people to share their knowledge, opinions, preferences, and skills far more efficiently. These social software tools open up a whole new set of possibilities for collaborative work modelled on self-organising systems.

21 Though this can be minimised with good search tools.
Social Software

…the slogan for the Footprints project is:

We all benefit from experience;
preferably someone else’s

Alan Wexelblat and Pattie Maes (1999, p. 271), emphasis in original

A few perfumes and pheromones aside, we humans seem noticeably lacking in native trail laying skills. Here the contemporary cyborg has a distinct edge, for she is already an electronically tagged agent. ...we can automatically lay electronic trials, which can be tracked, analyzed, and agglomerated with those laid by others.

Andy Clark (2003, p.145)

The electronic trails to which Andy Clark refers in the above quote is a large part of what makes social software possible. What others have done becomes known to current users in the form of stigmergic signals left in the environment. Each user benefits from the experience of the others.

In general, social software supports direct and indirect communication among a group of users. The focus of this discussion will be on indirect forms of communication, since they are typically asynchronous and stigmergic in nature. Much of the social software available is online, browser–based and freely available to anyone. As such, the groups that result can often number in the thousands. This creates a form of collaboration that is quantitatively and qualitatively different from traditional software designed for computer supported collaborative work/learning (CSCW/CSCL). Whereas these traditional forms of collaborative software support small, defined groups, social software typically supports what Dron and Anderson (2007) refer to as networks and collectives.

Some types of social software and their supported functions

Social software can refer to a very wide variety of online software (Dron, 2006; Owen, Grant, Sayers, & Facer, 2006) and includes:

- social networking sites like Facebook or MySpace
- wikis like Wikipedia
- social bookmarking and tagging sites like Delicious
- sites where users can post reviews of movies, music, products etc.
- commercial sites that support reviews and ratings, and provide purchase recommendations
• discussion forums (and blogs and news sites with discussion forums)

Many forms of social software incorporate multiple features and so may combine many or all of the above variants (see Table 3). User data can be collected passively (implicit) based on the actions the user performs for her own benefit – these data are a by-product of the users’ primary activities – or actively (explicit) (Recker, Walker, & Lawless, 2003; Wexelblat & Maes, 1999). In the latter case, the user is usually asked to perform some additional task, making some kind of judgement. Usually this involves providing an explicit rating, or categorising or reviewing something. In some cases, a user’s action might serve both the user’s immediate goals and take the form of an explicit judgement. For example, categorising a document with a quality indicator may be for one’s own benefit as well as being useful to others if shared. In general, asking users to add explicit information that does not directly serve the user’s current goal is less desirable. Providing the user with direct benefits for all actions generates data more quickly and avoids the “cold start” problem (Dron, 2005; Dron et al., 2001; Vassileva & Sun, 2007). A site of movie reviews and ratings by moviegoers is not very useful until it accumulates user data, for example. Encouraging an initial group of users to provide reviews and ratings can be difficult, however, since users will not want to use the site until it has some reviews that might assist them.
### Table 3: Examples of social software

<table>
<thead>
<tr>
<th>Site name</th>
<th>URL</th>
<th>Audience</th>
<th>Key functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia</td>
<td><a href="http://www.wikipedia.com">www.wikipedia.com</a></td>
<td>General</td>
<td>Anyone can edit any page in the encyclopaedia</td>
</tr>
</tbody>
</table>
| StumbleUpon| www.stumbleupon.com|                  | • Bookmarking  
• Tagging  
• Website reviews  
• Creating a social network |
| Delicious  | delicious.com      |                  | • Bookmarking  
• Tagging |
| Amazon     | www.amazon.ca      | Consumers        | • Book reviews and ratings  
• Book recommendations from the site to the user |
| Epinions   | www.epinions.com   |                  | • Product reviews and ratings |
| CiteULike* | www.citeulike.org  | Students and Academics | • Saving references  
• Downloading libraries to reference manager software (e.g., Endnote)  
• Tagging  
• Adding notes to references  
• Forming groups |


---

**Collaborative filtering and recommender systems**

One common type of social software is collaborative filters or recommender systems (Recker et al., 2003). They often work in similar ways, providing users a way of filtering a large set of options according to some category of interest or an algorithm is applied and direct recommendations are provided to the user. As an example of the latter, Amazon will e-mail customers book recommendations. These are based on the books the customer purchased in the past and new purchases made by other customers (who also bought some of the same books). This results in a recommendation in the form “People who bought X (like you did in the past) have also bought Y recently”.
Collaborative bookmarking and tagging

Another common type of social software is collaborative bookmarking and tagging systems\textsuperscript{22}. These features often go together, typically a website is bookmarked (and added to a database) and the user may then tag it\textsuperscript{23}. Tagging involves categorising the website using any number of keywords entered by the user. These tags may also be selected from a list of tags that the user has already used, or a set of tags created by the community of users. Since the set of community tags is often enormous only the more popular tags are typically displayed.

The community tags evolve in a decentralised fashion; there are no rules or screeners vetting the tags. These tags can also be used for browsing websites since the tags function as hyperlinks to a list of websites so tagged. Browsing by tags\textsuperscript{24} is like using keywords in a search, with the added benefit that the user knows that there are at least some pages associated with the tag. The tag also represents some user’s (or many users) interpretation of the website, not just a simple keyword text match. Tag clouds (see Figures 1 and 3 below) provide additional information to the user about the popularity of the tag by showing popular tags in larger font sizes.

\textsuperscript{22} Since most social software supports multiple functions, bookmarking and tagging might be better considered key functions of many social software systems.

\textsuperscript{23} Any digital object, such as photographs and videos, may also be bookmarked and tagged.

\textsuperscript{24} Some have referred to this as “pivot browsing” since clicking on tags provides the user with another viewpoint from which to see the library (see Bateman, 2007).
A very popular bookmarking site with a tagging feature is Delicious (http://delicious.com). On some websites the following icon appears (also see Figure 2; links to the bookmarking applications are shown in the red box). Clicking on it will allow you to upload a bookmark to Delicious and tag the bookmark as well (you must be a registered user of Delicious, which is free). The benefit of using Delicious is that the user’s bookmarks can be accessed from any network connected computer. As well the tags provide a more flexible means for categorising sites than using folders. A bookmarked website can be tagged with many tags, or categorised in many ways. Usually bookmarks are saved to just one folder, one category. Although users tag for their own benefit, a by-product of these efforts is that the Delicious website accumulates bookmarks and their associated tags from all of the users. The resulting collection of tags is often referred to as a “folksonomy” (Bateman, 2007; Owen et al., 2006) to highlight the fact that it is a kind of taxonomy created by a user community, not experts.
Figure 2: Bookmarking applications (see red box) displayed on a journal article website

Golder and Huberman (2006) have studied the pattern of the tags from the Delicious site\textsuperscript{25}. They analysed the proportion of tag use associated with particular URLs (i.e., bookmarks). Their analyses revealed that the proportion of tag use stabilised after about 100 bookmarks (i.e., 100 users bookmarked the URL and tagged it). This means that if the most popular tag was used about 20\% of the time after 100 users it was probably still being used about 20\% of the time after 200 users. This also means that, for any given URL, the rank order by popularity of the tags remains relatively fixed after 100 people have bookmarked the URL. This phenomenon is an emergent property of collective tagging. As Golder and Huberman indicate, this occurs because of imitation and shared knowledge. By seeing the tags used by others users may also choose the same tags. This is imitation, or exploitation, made possible by the community tags displayed by the Delicious interface. However, only a few of the most popular tags are shown to the user when they bookmark a URL. However, the stability observed in the proportions of use occurs for all tags, even relatively unpopular ones. For these unpopular tags, the observed stability must come from knowledge shared by the users, since they are unaware that others are using the tag. It also means that users do not merely imitate, or exploit, the

\textsuperscript{25} The tagging data were retrieved between June 23-27, 2005.
tags already generated but continually contribute based on their own knowledge. As a group then, there is still exploration, although the pre-existing shared knowledge across the users still often leads to stability.

CiteULike.

An especially useful site for scholarly work that supports tagging is CiteULike (http://www.citeulike.org), which is now sponsored by Springer publishing. CiteULike has many features; a subset of these will be discussed here. The primary purpose of CiteULike is to allow users to create an online reference library. When a user locates a reference – at a journal publisher’s website, for example – clicking one button uploads all reference information to the user’s online library. Tags can be associated with these references and can be displayed in a tag cloud (see Figure 3 below). The tags allow the user to quickly locate articles by any number of categories. As with Delicious, tags are also shared across all users. Clicking on a tag in social areas of the site (i.e., outside of the user’s personal library) will show all of the articles ever posted with that tag.

![CiteULike tag cloud](http://www.citeulike.org)

**Figure 3:** The author’s personal CiteULike tag cloud (September 1, 2008)

The social, and self-organising, aspect of CiteULike occurs through the user’s ability to learn from others by following the stigmergic signs deposited in the browsing environment. These stigmergic signs are the hyperlinked 1) tags, 2) names of the article authors, and 3) names of other users and groups. These hyperlinks are not designed by
anyone but get added to the website through the article contributions of the users themselves, unlike typical websites.

Tag popularity, which is visually indicated in the tag clouds, is a particularly important and emergent phenomenon. Users tag articles based on their own knowledge but are also likely to adopt tags common to the community. Adopting these common tags will allow the user to find articles posted by others; users will be using a common language of tags.

Learning about other ways (other frameworks) of interpreting the articles can also be important. Take for example, the posting of the article by Goldstone and Janssen (2005) to CiteULike (see Figure 4). The tags for this article provide useful information about how the seven users and six groups categorised the article. For instance, the tag “complexity” situates this article within the larger framework of complex systems although there is no specific mention of that concept in the abstract. Someone interested in complexity might find this article when browsing using the community tags, whereas a keyword search might not locate the article at all. The benefit of the tags is that they result from human judgement; a user’s semantic representation of the article based on his understanding of a field. Combining the knowledge of many users results in articles being categorised in many ways reflecting different disciplinary perspectives. This can permit scholars in separate academic fields to see relationships or linkages across fields of study.

---

26 This will generally be a sub-community of all the users in CiteULike. The exact membership of the community will be unknown to any user and will also be constantly changing.

27 Users often post articles to both their own library and a group library so in this case there may have been fewer than 13 separate postings.

28 Pirolli and Chi (Chi & Pirolli, 2006; Pirolli, 2007) have discussed this with respect to “brokerage” across fields.
Other social features of CiteULike include the ability to create groups. These groups can be of any size: a lab group, members of an academic society, etc. Any articles posted by one member are then shared by all, with any associated tags and notes. Users can also create a “watchlist”. When a user locates another user (or group) that shares her interests she can watch that other user’s postings. For example, students could watch the articles uploaded by their professors or colleagues. Finally, the ‘neighbours’ function returns a list of other users who have posted some of the same articles that you have. This is a rather direct means of findings other users who share your interests, who you may then watch.
Summary of Social Software

Although not explicitly designed using a self-organising systems perspective, many types of social software support multiple forms of indirect stigmergic interaction. They allow large numbers of people to form groups – explicit groups or loose associations formed through watchlists and shared tags, for instance – based on their interests and information needs. Many educational settings might also benefit from the use of social software. There are often hundreds of students in university courses, thousands taking the same courses in different universities and high schools, a multitude of continuing education students in various professions and trades, and life-long learners. This reality can provide social software systems a very large, diverse set of learners who collectively possess varied knowledge and experiences. They may thus be capable of augmenting learning resources with information (tags, notes, reviews, etc.) that is relevant to their learning goals. This possibility is even more compelling given that learners, almost by definition, are often in need of guidance and would be more likely to seek out and attend to useful information of this kind. A discussion of social software that has been designed for education follows.

Social Software for Education

This paper explores an alternative approach to the use of computers in education, where machines are not in control nor are they the tools of teachers, but instead amplify and embody the combined intelligence of the learners who use them. In such systems the machines knit together with their users to form a landscape, allowing emergent behaviours based on the values and knowledge of the communities that inhabit them to shape them. This is possible because computers occupy a unique position as the tools and the medium as well as the environment in which interactions between people occur.

Jon Dron (2007b, p. 201)

Scholars have begun developing social software that is designed for education (Bateman, Brooks, McCalla, & Brusilovsky, 2007; Brooks, Hansen, & Greer, 2006; Dron, 2007a, 2007b; Dron et al., 2001; Dron, Boyne, Mitchell, & Siviter, 2000; Farzan & Brusilovsky, 2005, 2006; Koper, 2004; Recker et al., 2003; Tattersall et al., 2005; Vassileva et al., 1999). In some cases, scholars have very explicitly designed their software using a complex or self-organising systems framework (Dron, 2007b; Dron et al., 2001; Dron et al., 2000; Koper, 2004; Tattersall et al., 2005).
These approaches share a few things in common. First, the systems are designed to allow many students to access and amend, add to, or augment learning resources. As discussed earlier the information students provide can be collected passively (e.g., time students spend using a resource) or actively (e.g., asking the students to provide a rating of the resource). Second, the learning resources or information about them (often called metadata) are in some way modified over time. This is based on the learners’ use of the resources and any information they added or attached to the resources. Metadata may indicate the target audience, difficulty, uses, or type of content – review, study, theoretical piece, etc. – of a learning resource. Metadata may be indicated using text describing the learning resource or by signs and symbols in the software environment (e.g., icons adjacent to name of the learning resource). Finally, as a direct result of the two features above, the resulting learning resources or metadata about them are not completely pre-designed by instructors or designers. Based on how many learners use a learning resource and what they do with it – tagging, annotations and note taking, etc. – stigmergic signs are attached to the resource. Learners affect the environment through their actions but the environment, in turn, influences these actions by presenting learners with the community’s current “opinion”. The form that learning resources take emerges over time as a result.

Most of the social software for education has been designed primarily to assist learners locate useful learning resources given their current goals. The resources generally remain the same (e.g., the texts themselves are not modified) but their associated metadata changes over time.

Two key motivations generally guide this work. First, while metadata is often added to learning resources by authors or designers this process is time consuming and costly (also see Bateman et al., 2007). This often results in poor or absent metadata. Additionally, there are many resources available online that learners may wish to consult. These resources were not intentionally created for educational purposes and typically possess no metadata, or metadata of the wrong kind. Second, there is some concern that even when authors add metadata it does not address the learners’ goals. Metadata that results from the actual learners’ activities potentially overcomes both problems. All learning resources receive metadata as a by-product of the learners’ use of the resources.
As well, the metadata that results should be more likely to address their actual goals. An examination of some of the social software developed for education is presented below (see Table 4 as well).

Table 4: Examples of social software for education

<table>
<thead>
<tr>
<th>Software</th>
<th>Sources</th>
<th>Data collected from or about the learners</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoFIND*</td>
<td>(Dron, 2002, 2005, 2007b; Dron et al., 2001; Dron et al., 2000)</td>
<td>• Bookmarks (adding resources)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Rating resources using Qualities</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Qualities and Topics can also be added by the learners</td>
</tr>
<tr>
<td>*Collaborative Filter in N Dimensions (<a href="http://jondron.cofind.net/">http://jondron.cofind.net/</a>)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Open University of the Netherlands Group</td>
<td>(Janssen et al., 2007; Koper, 2004; Tattersall et al., 2005)</td>
<td>• Lesson order followed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Lesson success</td>
</tr>
<tr>
<td>Altered Vista</td>
<td>(Recker et al., 2003; Walker, Recker, Lawless, &amp; Wilensky, 2004)</td>
<td>• Ratings of resources</td>
</tr>
<tr>
<td>AnnotatEd</td>
<td>(Farzan &amp; Brusilovsky, 2005, 2006)</td>
<td>• Annotations (highlights, notes, rating)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Pages visited</td>
</tr>
<tr>
<td>OATS**</td>
<td>(Bateman, 2007; Bateman et al., 2007; Bateman, Farzan, Brusilovsky, &amp; McCalla, 2006)</td>
<td>• Highlights</td>
</tr>
<tr>
<td>**Open Annotation and Tagging System (<a href="http://ihelp.usask.ca/OATS">http://ihelp.usask.ca/OATS</a>, demo only and source code)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CoFIND

CoFIND is a collaborative filter that allows learners to add resources to a community database (Dron, 2002, 2005, 2007b; Dron et al., 2001; Dron et al., 2000). It should be noted that unlike many other social software systems for education Dron has very explicitly designed CoFIND based on complex, self-organising systems. CoFIND was also intentionally designed to be used by a defined learning community with shared goals (e.g., students in a course). It has been developed and used exclusively in higher
education. Since CoFIND has been through many versions the following will be based on some core features.

When adding resources (i.e., bookmarking) learners may also assign the resources to a topic and associate qualities\(^{29}\) with them. Once a resource has been added to the database it is visible to all of the learners; these are not personal but community bookmarks. The topics are arranged hierarchically like a traditional categorisation scheme. Learners select a topic based on the content of the resource. The learners may also add new topics. Qualities are like tags except that they refer to the nature or quality of the resource (e.g., useful, complicated, good overview, easy to understand) and not the category to which the content belongs. Resources can be associated with multiple qualities. The learners may also add new qualities. The idea behind qualities is that they represent the multitude of ways in which learners, and perhaps teachers, assess learning resources. The acceptability of a resource depends on your current goals and is multidimensional, hence many qualities. An ‘easy to understand’ introduction to a topic is very useful when you are starting to learn about something, whereas an article with ‘rich content’ might be needed when writing the end of term paper.

To encourage the system to self-organise CoFIND is designed to calculate the popularity of resources, topics, and qualities. As such, the actions of the learners drive a kind of evolutionary process whereby any of these can fade away or gain prominence. For example, resources are displayed in an ordered list depending on the topic of interest and quality selected (using drop-down menus). Popular resources about that topic that also have been highly associated with that quality are placed at the top of the list. The learner is therefore given a type of recommendation since the resources are filtered and sorted based on the community’s judgement. Similarly, when bookmarking a new resource a list of qualities is displayed for the learner to choose from. The order of the list is based on the popularity of the qualities. As Dron explains, “both data and metadata are in constant evolutionary struggle with each other, competing for list position, which influences further success through a positive feedback loop” (2007b, p. 205).

\(^{29}\) The learner rates each resource across any number of qualities using a binary scheme (e.g., “it is/is not useful”)
Action research studies were conducted during the development of CoFIND (Dron, 2002), typically in intact educational settings (i.e., university courses). These studies mostly focused on whether the software was functioning to promote a self-organising system. These studies have not been published, and are of a qualitative and design focused nature. No studies were completed to assess learning outcomes as this was not the researcher’s interest at the time (Dron, 2002, p.256). One problem that was noted is called the “cold start” problem (Dron, 2002, 2005, 2007b). CoFIND needs enough data from the learners – resources and their associated topics and qualities – before any community consensus can be inferred. Until that time, general recommendations in the form of filtered and sorted lists will not be useful. It is difficult therefore to get enough learners to contribute early in the process since there are no incentives to use CoFIND.

CoFIND provides learners with information about learning resources and promotes stigmergic interaction. However, no explicit directions or recommendations are provided to the learner. List order focuses the learner’s attention on certain items but ultimately the learner may decide which resource or quality is best suited to her current task. A group at the Open University of the Netherlands has also designed systems explicitly based on a self-organising systems approach. But in this case, specific directive recommendations are made to the learner. Their work is described next.

Open University of the Netherlands Group

This group has been interested in social navigation systems. In particular, they have investigated systems that provide learners with recommendations about the order in which lessons should be completed (Janssen et al., 2007; Koper, 2004; Tattersall et al., 2005). Koper (2005) has also used agent-based modelling to perform simulations of learners using these systems.

In one study, 11 lessons about the Internet were created and made available to learners online for free (Janssen et al., 2007). Each lesson was designed to take an average of two hours to complete and learners who completed all of the lessons would receive a certificate. The study attracted 808 learners, 60% of whom completed a background questionnaire. Of those who responded, the learners tended to be over 45
years old (57%), with high educational attainment – professional or university level (63%), and poor computer skills (48%).

Lessons not yet completed were randomly ordered every time the list of lessons was viewed. Learners could complete the lessons in any order they chose. However, the experimental group used a self-organising version of the course. A recommendation about which lesson to complete next was presented to learners in this group after they successfully completed a lesson. A lesson was considered completed if the learner successfully answered three of five questions. The recommendation was based on the proportion of learners who went on to successfully complete another lesson. For example, if after completing a lesson 70% of the learners went on to successfully complete lesson 3 and 30% lesson 4, then the recommendation would have a chance 70% of being lesson 3. Random selection is used to prevent the learners from all following the same path before other possibilities have been considered. It prevents the system from entering a local optimum (Koper, 2005, paragraph 2.14)

Results indicated that more of the learners in the experimental group completed all of the lessons (40%) than the control group (33%) (Chi-square = 4.04, df = 1, p < 0.05). (Janssen et al., 2007, p. 790). This is a fairly weak effect, namely a 7% completion rate advantage for the experimental group. Given that the recommendations are self-organising and not designed by experts this is still somewhat impressive. However, the software was not compared to a prescribed lesson order from an instructor or designer. Rather, both groups were presented with the lessons in a new random order every time they viewed them. The study would have been more informative had there been an additional control group following a prescribed order.

Additionally, the criterion for successful completion of a lesson was quite low (3 of 5 correct answers). A more selective evaluation might improve the overall performance of the system by ensuring the recommended path was based on learners who had truly learned the material presented (of the recommended lesson). A weighting scheme could also be employed, whereby learners who performed better on subsequent lessons would influence the recommendation more than others.

30 “[this is done to] prevent sub-optimal convergence by allowing randomness in the proportion of learners who follow the ‘pheromone-based’ advice provided” (Koper, 2005, paragraph 2.14).
31 Incorrectly reported as df = 2 in the article.
 Whereas this approach used passively collected information – about the paths learners took and their success with lessons in that path – other systems have been based on actively collected information. Learners are asked to provide explicit ratings about learning resources that are then used to make recommendations. Altered Vista is one such system and is described next.

**Altered Vista**

Altered Vista is a collaborative filter that works based on learners providing explicit ratings of learning resources (Recker et al., 2003; Walker et al., 2004). In three studies reported in Recker et al. (2003) the information provided by the learners was different in each case. The focus of this brief review will be on the third study for which one overall rating of each learning resource was provided using a 5-point Likert scale. The learner was asked to rate the usefulness of the resource given their learning task. A learner’s set of ratings is then used to make recommendations to her about other useful learning resources. Recommendations are based on finding learners that have rated resources similarly; a learner’s “neighbours”. New resources are recommended based on what these neighbours might suggest.

The third study involved 375 undergraduate students in a family sciences course. Although four groups were created the reported results focused only on the two groups that received recommendations. Forty web resources were included for each of five case studies, for a total of 200 resources. Students read a case study and then read resources that might be useful with respect to that case.

For the first three cases students were provided with eight resources chosen at random and were asked to rate all of them. For the last two case studies the first group (n=89) received recommendations based on a random selection while the second group (n=87) received personalised recommendations from their neighbours. Participants in both groups received approximately three recommended resources on average. Participants read and rated these resources. All participants also completed a post-task survey and a unit exam that included an essay question. Although the group that received
personalised recommendations found them to be more useful (82%)\textsuperscript{32} compared with the group that received random recommendations (32%) there were no significant differences regarding the essays from the exam.

Many social software systems do not employ direct recommendations but rather attach signs to resources that indicate how much the community has used them. The learner is then able to select resources, in part, by using this information. AnnotatEd is such a system and is described next.

\textit{AnnotatEd}

AnnotatEd is a social navigation system (Farzan & Brusilovsky, 2005, 2006) that collects two types of information about the learner. First, it tracks which pages a learner goes to and how long the learner spends reading the page. Second, it records the annotations that learners add to a page. Learners are able to highlight sections of text, add notes to a page, and indicate if they liked the page (by clicking on a check box labelled ‘Praise’). Using these sources of information AnnotatEd places icons adjacent to hyperlinks that provide information about the community’s use of the linked page. Because of the different sources of information, Farzan and Brusilovsky (2005; 2006) refer to these as traffic-based and annotation-based social navigation support. One icon indicates the traffic to the page – based on the number of times the link has been visited. The other icon indicates the number of annotations made on the page. The centre of each icon displays the learner’s personal activity on the page as well. For both personal and community information, colour saturation is used such that darker colours indicate more activity (higher traffic or more annotations). A third icon was also incorporated in later versions to indicate the number of learners that liked the page (i.e., those that had clicked ‘Praise’).

A study was conducted using AnnotatEd in an undergraduate introductory programming course and a graduate level information retrieval course (Farzan & Brusilovsky, 2006). Both classes had 12 students who used the system throughout the semester to access resources relevant to the course. Results suggest that the annotated pages were visited more often (i.e., higher traffic). The extent to which traffic to the pages

\textsuperscript{32} Based on the percentage that responded 4 or 5 on a 5-point scale to the question “Altered Vista provided
is caused by the annotations is not clear, however. Indeed, traffic to the pages that were eventually annotated was still higher compared to pages that were never annotated at all, even before the former were annotated for the first time. This indicates that learners were going to these pages based on either the information in the hypertext link itself or the traffic icon, since annotations had not yet been added.

Learners can add highlights and notes to web pages using AnnotatEd. This information is primarily used to adjust the annotation icon, signalling the extent to which the community has annotated the page. Other systems have placed more of an emphasis on the within-document annotations as signs in themselves. OATS is such a system and is described next.

**OATS (Open Annotation and Tagging System)**

Unlike other social software which typically involves document-level metadata (e.g., tags associated with entire documents) OATS supports “within document”, or local metadata (Bateman, 2007; Bateman et al., 2007; Bateman et al., 2006). This is accomplished by allowing users to highlight sections of text. Highlighted sections may also be tagged and have notes attached to them. In addition, when a user selects an area of text to highlight a list of the ten most popular tags applied to the document by the community is presented to the user. A user may also turn on and view “community highlights” which are generated based on the overlap across all users and is displayed using colour saturation (i.e., a progressively darker background colour indicates greater overlap). These two features, knowledge of the popular tags and which sections of the text the community has highlighted (and possibly tagged or made notes on) provide the user with stigmergic signals.

A small case study was completed with 12 senior computer science undergraduate students (Bateman, 2007). Students read three articles using OATS over two weeks as part of a course. There were ten web pages of text in total (for all three articles...
It should be noted that the “community highlights” function defaults to off whenever a new page is viewed.

The results suggest that the students were motivated to view the community’s highlights as they turned on this feature approximately three times per page (statistics per student, for all three articles, were $M = 29$, $SD = 18$, $min = 7$, $max = 63$). There was a lot of variability in the use of this function (and the distribution is positively skewed), so claims about what the mean suggests are somewhat misleading. However, all of the students used this function with the lowest usage being seven, which is just shy of once per page. As such, the data suggest that students turn this function on quite often to examine where the community has been focusing their attention (i.e., highlighting and tagging).

These are tentative results, given the sample size and the large variability among the students. Indeed, one student accounted for 40% of the tags applied. A larger study would be beneficial since individual students would not affect the descriptive statistics as much, and clear outliers could be removed if necessary (by using a trimmed mean, for instance).

**Summary of Social Software for Education**

Evidence that these systems generate self-organising metadata with educationally useful properties is limited at this stage. These systems are relatively new and the research community is still quite small. Most efforts have been on conceiving and building the software. Although some empirical studies have been performed many have suffered from small samples, limited performance measures, or insufficient controls to address questions about the direction of causal relationships. Also, measures of system and student performance vary between studies and makes forming general conclusions difficult.

These systems attempt to create new forms of learner interaction and ultimately new ways of learning. Comparisons between these approaches and older ones may be misguided if the outcomes measured are framed with respect to one approach or the other. Just as observations in science can be “theory-laden”, evaluations in education can also be “theory-laden”. For example, when metadata is student generated valid questions include: “what metadata emerged?”, “when and how did this take place?”, and “what was the
quality of the metadata?” When an author designs the metadata, such questions are either not relevant or simply not considered.

When can we say that social software used in an educational context worked well? There are at least three levels at which this can be answered:

1. Did the group’s activity result in some stable set of metadata that can be described?
2. What was the quality of this metadata relative to the learners’ goals or tasks?
3. Was there any noticeable improvement in the quality of the learners’ work, knowledge, or skills from having used the software?

Although easily stated, these questions are more difficult to answer than it appears. Because of the nature of social software the metadata is always changing and the information one learner interacts with is not the same as the next learner. Moreover, the learners are by definition interdependent since they interact through stigmergic signals. These two features – dynamic learning resources and interdependent learners – means that social software does not lend itself to traditional research designs. For example, inferential statistical analyses are not appropriate when the units of analysis are not independent. Since learners using social software by definition influence each other, many of the statistical results reported above are questionable.

There are ways around these difficulties. For instance, one might use an initial group of learners to generate the metadata while another set of learners uses it afterwards. For the second set of learners, the resources are stable and the learners are independent of each other. This is not an optimal solution since in actual educational use no such separation would occur. Learners would use the software, dynamically generating metadata while completing their learning tasks. However, for providing initial proof of effectiveness this type of two-part study may be a useful approach.

Traditionally, metadata has been provided by the authors of documents. This is very similar to the use of text signals by authors. In both cases, authors provide information to learners (or readers) that is designed to guide or assist them. A definition of text signals, the types of text signals used by authors, a framework for understanding text signals, and a review of the research on text signals is presented next.
Text Signalling

Text signals are writing devices that emphasize aspects of a text’s content or structure without adding to the content of the text.


There is a fairly extensive body of research on text signalling, with much of the more recent work by Lorch and his colleagues (Lemarie, Lorch, Eyrolle, & Virbel, 2008; Lorch, 1989; Lorch & Lorch, 1996; Lorch, Lorch, & Klusewitz, 1995; Lorch, Lorch, Ritchey, McGovern, & Coleman, 2001; Mautone & Mayer, 2001; Meyer & Poon, 2001; Naumann, Richter, Flender, Christmann, & Groeben, 2007). Text signals, as the quote above indicates, are writing devices that emphasise parts of the text or indicate the structure of the text. They may include a variety of devices: previews, overviews, summaries, titles, subheadings, typographical cues indicating importance (e.g., bold, italics), and even phrases that emphasise content or explicitly describe structure (e.g., “in summary”, “first of four parts”).

More recently, Lemarie et al. (2008) have proposed a new model, and definition, of text signalling based on a “Textual Architecture Model”. In this framework, a text is transformed into a prototext. This prototext specifies both the content of the text (sentences) and metasentences that refer to the content sentences. These metasentences “express an author’s intention to perform a textual act” (Lemarie et al., 2008, p.31); which is derived from Speech Act Theory (Searle, 1969). Lemarie and colleagues propose that signalling is a text act designed by the author to have a desired effect on the reader. The actual text is created by transforming the metasentences into text signals.

A signal is the realization in a printed text of a metasentence, or set of metasentences, from the underlying prototext.

Lemarie at al. (2008, p.31)

As well, any given metasentence can be realised in the text with a variety of text signals. Since the author’s intention is to have an effect on the reader, the metasentences, and signals used to realise them, are treated as processing instructions to the reader from the author.
For example, typographical contrast expresses the author’s intention to emphasize particular text content. As another example, a system of headings communicates the author’s organization of a text. Thus, as illocutory acts, signals may be viewed as the realization in a printed text of an author’s instructions regarding how the text is structured and how emphasis is to be distributed across the text content. Finally, an instruction/signal by the author may be heeded by the reader with the result that the reader’s processing may be influenced (i.e., the perlocutory effect of the underlying metasentence). In short, identifying signals with metasentences provides a means to analyze the nature of the author’s “instruction” to the reader.

Lemarie et al. (2008, p.32)

Lemarie et al. (2008) characterise signals as having one or more of six functions (illocutory acts) and are either discursively or visually realised in the actual text. Additionally, they can be described by their scope of influence and location relative to the text they signal (see Table 5 below).

Table 5: Characteristics of text signals

<table>
<thead>
<tr>
<th>Text Acts – Functions</th>
<th>1. Demarcate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2. Organise</td>
</tr>
<tr>
<td></td>
<td>3. Label</td>
</tr>
<tr>
<td></td>
<td>4. Identify function</td>
</tr>
<tr>
<td></td>
<td>5. Identify topic</td>
</tr>
<tr>
<td></td>
<td>6. Emphasise</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Basic types</th>
<th>Discursive (written phrases)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Visual traces (spacing, typeface changes, etc.)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scope</th>
<th>Local or global</th>
</tr>
</thead>
</table>

| Location | Coincident, before, interspersed, or after the content signalled |

* Note that these characteristics are not always independent. For instance, typographical text signals (e.g., **bold**, *italics*, font **colour**) are necessarily both local in scope and coincident in location.

**Empirical studies of text signalling**

Most studies have shown that text signals produce better performance on the recall of the signalled content. A comprehensive review by Lorch (1989) found memory was improved for signalled content for a variety of text signalling devices (see p. 229). Since the study conducted (see Chapter 3: The Current Study) employed typographical text signals a brief review of the empirical findings related to these signals follows.

Typographical text signals are relatively simple. They are necessarily local and coincident with the content they signal. They are visually realised using typeface changes
and “expresses the author’s intention to emphasize particular text content” (Lemarie et al., 2008, p.32). Since locating the most relevant information within an expository text is a crucial task for successful comprehension (A. L. Brown & Day, 1983; Dole, Duffy, Roehler, & Pearson, 1991; Winograd, 1984) the use of typographical text signals that emphasise the most relevant content should often prove beneficial.

Lorch concluded that typographical signals improve memory for the signalled content and that memory for unsignalled content is generally unaffected, or sometimes inhibited (1989, p. 224). Most of the studies had used memory measures to compare performance on the signalled versus the unsignalled content.

A later study on typographical text signals by Lorch et al. (1995) also confirmed these findings. This study consisted of two experiments that used the same text. The characteristics of the participants, the text used, the experimental conditions, and the recall test used in both studies are presented in Table 6 below.

Table 6: Characteristics of the participants, the text, and the reading and recall conditions (Lorch et al., 1995)

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Participants</th>
<th>Text</th>
<th>Conditions</th>
<th>Recall test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Undergraduates in psychology courses N = 124</td>
<td>Expository • 2400 words • 28 target sentences of which 14 were signalled in the signalling conditions</td>
<td>None, Low and High signalling • Low, 14 of 28 targets signalled (5% of the words) • High, 50% of words signalled, this included the 14 targets and half of the other sentences</td>
<td>18 short answer questions covering content for all 28 target sentences immediate recall following reading</td>
</tr>
<tr>
<td>2</td>
<td>Undergraduates in psychology courses N = 80</td>
<td>As above</td>
<td>None versus Low signalling</td>
<td>as above</td>
</tr>
</tbody>
</table>

*Note: In the first experiment the text was read on paper whereas in the second experiment it was read on a computer monitor, one sentence at a time, to facilitate recording the time spent reading each sentence.*
From the first experiment it was determined that the signals were effective only in the low signal condition; the effect size was a 12% increase in recall for the signalled targets (43%) versus the unsignalled targets (31%). There were no differences across the conditions for the 14 targets that were not signalled (26%, 31%, and 26% for none, low and high signals conditions, respectively). The high signals condition produced similar recall for signalled (28%) and unsignalled targets (26%).

The second experiment in the study produced a smaller effect size for the low signal condition (i.e., a 6% advantage for the signalled content) but was consistent with the first. Also, it provided evidence that the signalling caused readers to slow down, indicating that the signalling may have affected the processing of the text. This suggests possible mechanisms to explain the improved memory, namely increased rehearsal of the signalled text or possibly a “deeper” form of processing. Perhaps readers spent the extra time forming coherence links between the signalled text and items in working memory, for example.

Typographical text signals seem to increase students’ immediate recall of signalled content without affecting memory for unsignalled content. This improvement appears to result from an increase in the time spent processing the signalled content. There may also be a difference in the strategy used to process the content. It should be noted that typographical text signals seem to be ineffective if too much text is signalled. For low signalling conditions, the effect size in the study by Lorch et al. (1995) using immediate recall and short answer questions was modest; a 12% and 6% improvement for the two experiments, respectively. Although beneficial, adding typographical signals to a text may require more effort than can be justified by the results they provide.

Adding signals to texts in practice

Although authors routinely use text signals, the systematic addition of text signals to educational texts to assist specific types of learners would prove to be time-consuming and costly. This is especially true for texts not initially written for a student population. In such cases, instructors would be required to modify the text. Hard copy texts would need to be digitised and manually signalled. This is unlikely to be done given the costs
involved, the small although beneficial effects from text signalling (small effect sizes)\(^{36}\), and issues of copyright. Essentially, such solutions do not scale well because of the large number of texts that would need to be modified manually\(^{37}\). As well, it may be difficult to anticipate the needs of learners without empirical data. For instance, key concepts in a text that are already well understood by a reader need not be signalled. However, supporting concepts that help clarify the main concepts may need to be signalled for readers who have very little experience with the topic.

Whereas text signalling is an act by the author designed to affect the reader, readers themselves annotate texts in ways “designed” to affect their current or future processing of the text. Additionally, if such annotations and marginalia (i.e., notes, highlighting or underlining, emphasis marks like stars, etc.) could be shared with others a new form of text signalling, or shared annotations (Marshall, 1998; Marshall & Brush, 2004), might be possible. One where the illucutory acts are from one reader, or set of readers, to yet other readers: processing instructions “by readers for readers”.

Digital text affords many opportunities for annotation, both individual and social. Since digital texts also permit annotations to be easily shared they afford interaction that is not normally possible with hardcopy texts (Marshall, 2005; Marshall & Brush, 2004). However, for digital annotations to become a viable option for readers, a portable reading device with annotation capabilities is needed (Schilit, Price, Golovchinsky, Tanaka, & Marshall, 1999). A brief review of a few such devices follows.

There are currently several e-book reading devices on the market (Table 7) which use electronic ink (e-ink) and can display fairly high resolution text and pictures in grayscale (4, 8 or 16 levels). The Irex Iliad (http://www.irextechnologies.com/), for example, uses e-ink and has no backlight: together these permit the device to remain functional for many hours on a single battery charge. This device also has a stylus for taking notes on the digital pages you are currently reading. Finally, the device has wireless capabilities. This device provides a unique platform for designing social software for reading. Students could use such a device with appropriate social software to read and

\(^{35}\) Signalling in general produces modest effect sizes (see Pressley and McCormick, 1995, p. 484)

\(^{36}\) A cost-benefit analysis probably makes this type of manual text signalling unjustified.
annotate texts, with the possibility that their underlines, annotations, summaries, questions, etc. could be collectively shared. Essentially a group – the whole class, a self-selected group of students, or multiple classes – could read and “work on” the text together.

Table 7: E-book reading devices*

<table>
<thead>
<tr>
<th>Manufacturer</th>
<th>Name</th>
<th>website</th>
<th>Specs/Features</th>
<th>Wireless?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon</td>
<td>Amazon Kindle</td>
<td><a href="http://www.amazon.com">www.amazon.com</a></td>
<td>1. 6 inch, e-ink screen</td>
<td>yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. 320 g</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. Bookmarking &amp; highlighting (via keyboard)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4. Text search</td>
<td></td>
</tr>
<tr>
<td>Irex</td>
<td>Iliad 2nd edition</td>
<td><a href="http://www.irextech">http://www.irextech</a></td>
<td>1. 8.1 inch, e-ink screen</td>
<td>yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>nologies.com/</td>
<td>2. 389 g</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. stylus: hand-written notes &amp; annotations</td>
<td></td>
</tr>
<tr>
<td>Sony</td>
<td>E-book reader</td>
<td><a href="http://www.sonysty">http://www.sonysty</a></td>
<td>1. 6 inch, e-ink screen</td>
<td>no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>le.com</td>
<td>2. 250 g</td>
<td></td>
</tr>
</tbody>
</table>

* Also see http://en.wikipedia.org/wiki/E-book_device

Summary of Text Signalling

Text signalling provides a means to deliver processing instructions to the reader. Typically these instructions are from the author to the reader. Evidence has shown that they generally improve memory for the signalled information. However, the beneficial effects of adding signals (such as typographical text signals) are relatively small. The cost of adding them to a text manually may often be high. As such, an efficient means of augmenting texts with signals is needed. Software to support the creation of collaborative text signals by the readers themselves was designed (see Chapter 2: Designing CoREAD below). However, before describing the software, a framework for integrating complex systems, social software and text signalling is needed. Distributed cognition is proposed as a useful framework for this purpose and is discussed next.

37 Similar arguments have been made regarding the manual creation of metadata for webpages (Bateman et al., 2007) and concept-based navigation support in adaptive hypermedia (Farzan & Brusilovsky, 2005).
Distributed Cognition

I propose a broader notion of cognition because I want to preserve a concept of cognition as computation, and I want the sort of computation that cognition is to be as applicable to events that involve the interactions of humans with artifacts and with other humans as it is to events that are entirely internal to individual persons.

Edwin Hutchins (1995a, p.118)

In this paper, I will attempt to show that the classical cognitive science approach can be applied with little modification to a unit of analysis that is larger than a person.

Edwin Hutchins (1995b, p. 266)

Distributed cognition was introduced around 1990 as a way of broadening the unit of analysis in cognitive science beyond the individual person (Hollan et al., 2000; Hutchins, 1995a, 1995b; Kirsh, 1999, 2006; Kirsh & Maglio, 1994; Pea, 1993; Salomon, 1993; Zhang, 1997, 1998; Zhang & Norman, 1994). The above quotes by Hutchins indicates that he saw this as more-or-less a direct “expansion” of traditional cognitive science (i.e., the information processing perspective) to a larger unit of analysis.

However, the spirit of traditional information processing focused on the individual’s internal symbolic processing of a situation. Less focus was placed on group interaction, the use of external artifacts, or the role of “culture” more generally (Kirsh, 1999). The extent to which distributed cognition and related frameworks – situated cognition (J. S. Brown, Collins, & Duguid, 1989), embodied cognition (M. L. Anderson, 2003), extended mind (Clark, 2001; Clark & Chalmers, 1998), and others (see Walmsley, 2008) – constitute an actual break with information processing may in fact be more serious than Hutchins suggested. The differences between situated cognition and information processing (i.e., the cognitive perspective) has been hotly debated, for instance (J. R. Anderson, Greeno, Reder, & Simon, 2000; J. R. Anderson, Reder, & Simon, 1996, 1997; Greeno, 1997, 1998). In any case, these other approaches bring into focus issues that were typically left out of traditional information processing. Left out need not imply dismissed as irrelevant, however. For example, Newell proposed levels of cognitive activity

---

38 It seems to have been conceived in the mid-80’s though the more prominent works appeared in the 1990’s.

39 Pea (1993) referred to it as distributed intelligence.

40 Kirsh (1999) in fact refers to “the spirit of the problem space approach” associated with Newell and Simon (1972), which is grounded in the larger information processing perspective.
organised roughly by the time needed for their completion (Newell, 1990); higher levels span hours, days, weeks, and so on. Such activity must clearly include interactions among people, and between people and external artifacts (at least those created at earlier times by the same person). It certainly suggests that a unit of analysis larger than an individual might be useful or necessary when addressing activities that occur over greater time periods.

The following presents the distributed cognition framework and explains what is meant by a larger unit of analysis. Several research studies are then described. It is then argued that a complex (self-organising) systems approach to distributed cognition is viable, particularly with respect to stigmergic interaction. Although this link was not typically addressed in the early literature on distributed cognition, there are some examples of more recent work that have done so (Gureckis & Goldstone, 2006; Heylighen, Heath, & Van Overwalle, 2004; Poirier & Chicoisne, 2006). Last, I will describe how social software can also be seen as a form of distributed cognition.

What is distributed cognition?

The principle idea of distributed cognition is that the unit of analysis extends beyond the individual. This means that to properly study cognitive activities one must examine the interactions an individual has with external resources and other people. In fact, it means treating the individual as a part of a larger cognitive system. Hollan at al. (2000) propose that there are three ways in which a cognitive task may be distributed (p. 176):

1. Distributed across members of a social group. Social organisation can then be seen as a kind of cognitive architecture. Information enters the social network and is processed as it is transformed and passed between individuals. In connectionist terms, the people themselves are the nodes in the social network.
2. Distributed between internal (mental computation, biological memory) and external (material and environmental) structures, or resources. Because of the role of external artifacts this implies embodied cognition as well (M. L. Anderson, 2003) since these artifacts must be perceived and manipulated in the course of the activity.
3. Distributed through time such that earlier events affect later ones. As Kirsh puts it, “history matters” (2006, p.250).

All of the above types of distribution lead to a serious incorporation of culture, where culture might be thought of as partial solutions to problems manifested as tangible artifacts, intangible processes, learned “heuristics”, and social interactions and conventions that were created in the past but affect the present activity.

At the heart of this linkage of cognition with culture lies the notion that the environment people are embedded in is, among other things, a reservoir of resources for learning, problem solving, and reasoning. Culture is a process that accumulates partial solutions to frequently encountered problems. Without this residue of previous activity, we would all have to find solutions from scratch. We could not build on the success of others.

James Hollan, Edwin Hutchins and David Kirsh (2000, p.178)

Not amplification

Although distributing cognition among external and internal resources, across people, and over time often leads to improved performance this should not be considered a kind of cognitive amplification. Rather, the cognitive activity has changed – a computation in the head has been replaced by a perceptual judgement, for instance – and so no amplification of the original process(es) has occurred (Hutchins, 1995a, p. 153-155). This is important as it affects our judgement about who (or what) performed the cognitive task. It is simply not the case that the individual performed the same processes and her inherent cognitive ability was simply amplified by a resource – another person or a cultural tool. Rather, a larger unit of analysis has completed a series of processes leading to the completion of the cognitive task. The processes completed “in-the-head” of the individual are simply a part of this whole. Moreover, these are often different processes than those that would have been performed by the individual “alone”, that is, without distribution of some sort.

In order to appreciate the distributed cognition perspective more fully, some research studies will be described next.

Research studies based on the distributed cognition perspective

The following research studies are grouped by the general methodology used. These methodologies include ethnography, controlled experiments, and approaches based on a
design perspective. Last, a proposed methodology (Hollan et al., 2000) based on using digital technologies for data capture is also described. This proposed methodology is a natural fit with social software.

**Ethnographic studies**

The work by Hutchins on ship navigation systems and airline cockpit systems (1995a; 1995b) was conducted using ethnographic methods. In such an approach, the researcher studies pre-existing cultural practices. Careful observations – recorded using audio or video, and journal notes – were conducted of the collaboration among several participants using various task-specific artifacts. The goal of this research approach is to uncover the way in which a small group of individuals and their tools (i.e., the artifacts) interact to solve specific problems. The unit of analysis is the navigation system or cockpit system (i.e., the crew and artifacts combined), respectively.

**Navigation system.**

Perhaps one of the most well known works on distributed cognition is Hutchins’ book *Cognition in the Wild* (Hutchins, 1995a). What is presented here focuses on the navigation task known as Sea and Anchor Piloting Detail conducted by a military naval crew, as described in Chapter 3. This task is performed whenever the ship, a large amphibious helicopter transport, enters a port or is in hazardous waters. It involves plotting a safe course, and then noting the ship’s actual position on a chart (a map) in real time by visually sighting landmarks. Several related sub-tasks will be discussed to examine how the larger unit of analysis – the navigation system – represents the world and works on this representation to solves problems.

Key artifacts of the navigation system are the charts (maps) used to represent the world. These charts could be considered the best example of the navigation system’s internal representation of the world. To know where in the world the ship is, the navigation system adds and transforms information on the current chart being used. Ultimately, the position of the ship is read off the chart.

For the Sea and Anchor Piloting Detail a series of charts of the harbour in question would be used. These would be prepared ahead of time and stacked in the correct order to allow new charts to be revealed as needed. More interestingly, navigation crews augment
the chart(s) for their home harbour. This includes marking the chart to visually identify shallow waters, which is specific to the ship because each ship has a different draft (depth). As well, new landmarks may be added over time as new buildings are erected. Finally, the entry and exits tracks used by the ship are often permanently added to the chart in ink. This process of externalising simple judgements (e.g., is the depth too shallow?) and the results of more complex tasks (e.g., deciding the ship’s entry track to the harbour) by placing structure in the chart changes the future work that individuals within the navigation system will perform.

When performing the Sea and Anchor Piloting Detail the navigation system must determine the ship’s position in the world, its speed, and make recommendations to the bridge about heading and speed changes. What follows is an examination of how the problems of determining location and speed are solved.

To determine the location of the ship in the world a mark is made on the chart. The sequence of events that leads to this is as follows:

1. An appropriate landmark is selected from the chart based on where the navigation system estimates the ship to be.
2. A sailor (a pelorus operator) visually locates the landmark using an optical instrument called an alidade and reads off the compass bearing displayed in the sight.
3. This bearing angle is recorded in a log in the charthouse and is then represented on a protractor (called a hoey).
4. A line is drawn on the chart (a line of position) originating from the landmark using the hoey.
5. The sequence is repeated for another landmark, usually on the other side of the ship.
6. Where the two lines on the chart intersect marks the ship’s position.

Several interesting observations should be noted. First, there is no need for any member of the crew to have an internal representation of where the ship is; at least not one that could properly solve the problem. Rather, the navigation system has the representation of interest. The solution to the problem is found through a transformation of the inputs – a series of re-representations – of the direction of the landmarks as seen by the pelorus.
operators. Ultimately, these transformations lead to new marks on the key internal representation, the chart. One might consider the chart the navigation system’s “mind’s-eye”. Second, no individual member solves the problem. Each member solves a sub-problem such as locating the landmark or drawing the line of position on the chart. Each member serves as a component of a larger cognitive system, the navigation system.

Another problem the navigation system must solve is calculating the ship’s speed. Although there are various tools for doing this Hutchins’ emphasises one particular means of doing so because it is an example of how cultural knowledge is not always stored in tangible artifacts. Determining the ship’s position is typically conducted every three minutes. This practice is not arbitrary. Rather, it is followed because doing so allows the crew to read off the ship’s speed directly from the chart using the ‘3-minute rule’. The math is quite simple. Three minutes is 1/20 of an hour and 100 yards is 1/20 of a mile. Measuring the distance in yards on the chart and removing the last two zeroes (dividing by 100) will give the ship’s speed in nautical miles per hour (knots). Simply put, 1500 yards on the chart means 15 knots, for instance. Here the practice of fixing the ship’s position every three minutes enables speed to be computed quickly and with few errors. The calculation of speed has been distributed but it has not been fully implemented in a tangible external artifact. Rather, the distribution involves the chart, tools for measuring distance on the chart, a social practice (take fixes every three minutes), and a “computation” rule that is internalised by individual crewmembers.

*Cockpit system.*

Hutchins (1995b) also employed his ethnographic approach to the study of a commercial airline cockpit. The title, “How a cockpit remembers it speeds”, was used to emphasise that the cockpit system – two pilots in combination with social practices and external artifacts (e.g., the airspeed indicators) – remembers important landing speeds, not the individual pilots. Not only is memory off-loaded to external media, Hutchins shows that the mental computations required of the pilots actually changes.

Unlike small aircraft, commercial airliners can change the shape of their wings in order to maintain lift during landing when the speed of the aircraft is reduced (reducing speed also reduces lift). The task studied, therefore, was the problem of when to extend
the flaps and slats to different positions during the landing sequence. This procedure begins with one of the two pilots (the pilot not flying) locating a card with flap/slat settings and matching airspeeds. Each aircraft is supplied with a set of such cards for various weights of the aircraft. To select the correct card from the set the pilot not flying reads the aircraft’s weight, displayed on a cockpit instrument. The card is then placed in a location where both pilots can read it. At this point we can see that the task has several artifacts that either store computations or decisions (the speed cards) or ‘sense’ the conditions of the world (the weight indicator). The pilots need not calculate the landing speeds since there is an external memory store of this information.

Next the pilots transfer the landing speeds directly to the airspeed indicator. Movable “speed bugs” on the outside of the airspeed indicator allow each pilot to represent the critical speeds, including the final landing speed, around the dial of the airspeed indicator. During descent the pilots do not have to represent the aircraft’s current speed or the critical landing speeds – when flaps and slats will need to be extended – as numbers. Rather, the pilots use the position of the airspeed needle in relation to the speed bugs to “see” where in the sequence they are. This is an example of how placing information in the environment sometimes permits people to do different mental computations that are less costly or less error-prone. The result is thinking differently not merely remembering less. It is important to emphasise that the pilots’ memories were not amplified. The artifacts did not change their capacity to remember but changed the actual process that they engaged in. A quantitative comparison task (comparing speeds) was exchanged for a perceptual location task (where is the airspeed needle?). As well, like the chart discussed in the previous study, the airspeed indicator with the set speed bugs serves as a key representation internal to the cockpit system, but not internal to any person.

These ethnographic studies show that over time cultural artifacts and practices evolve in order to transform difficult mental computations into more efficient, less error-prone judgements (e.g., seeing important airspeed spaces on a dial) or simpler mental computations (e.g., the 3-minute rule). Additionally, it is useful to treat the larger persons-practices-artifacts system as the unit of analysis.
Experiments

Unlike Hutchins, who studied intact cultural practices Zhang has conducted studies that experimentally manipulate the location of representations, either internal to the participant or located externally. An abstract task representation is divided into internal and external representations that may overlap or not. Specifically, the rules describing simple game-like tasks such as the Tower of Hanoi (Zhang, 1998; Zhang & Norman, 1994) and Tic-Tac-Toe (Zhang, 1997) were experimentally manipulated so as to be known internally (i.e., in a participant’s memory) or represented externally. The rules were then distributed between a human participant (internal representation) and an external representation (Zhang, 1997; Zhang & Norman, 1994) or between two participants (Zhang, 1998).

Externalising representations.

In three related experiments a modified Tower of Hanoi task – with larger disks on the top, and smaller on the bottom – was used (Zhang & Norman, 1994). The discussion below will focus on the results of the second experiment. Three rules governed the task: 1) only one object can be moved at a time, 2) an object can be moved to a new location only if it is larger than the objects already there, and 3) only the largest object can be removed from a location. Three isomorphic representations of the task were created. In two of the representations possible actions were constrained. In this way, some of the task rules (e.g., rule #2, only larger objects can be placed on top of smaller ones) were embedded in the representation. For example, in one isomorph different size cups of coffee were used. The cups were stacked one on top of the other. This prevented smaller cups from being place on top of larger ones, as they would fall through the larger cup. The result was that for one representation all of the rules had to be remembered, for the second only two of the rules (one rule externalised), and the third only one rule (two rules externalised).

Each participant completed the task once using each type of representation. The results showed that externalising one or two of the rules improved performance (greater speed, fewer steps, fewer errors) compared to remembering (internally) all of the rules. Note that the representations do not in any way make the series of operations needed to
reach the goal state clear. Rather, externalising the rules merely seems to allow the participants to better focus their attention and utilise their working memory resources for solving the problem.

The finding that the affordances and constraints of external representations can have dramatic effects on human performance is quite well known (Norman, 2002; Vicente, 2004). More interesting is Zhang’s study of the distribution of rules between two persons who are cooperating to solve the same problem, examined below.

**A trade-off between diversity and communication costs.**

In this study (Zhang, 1998), two participants (dyads) solved a variant of the Tower of Hanoi problem together but could not speak to each other, except to indicate a violation of a rule. Additionally, which of the rules each participant knew was manipulated. In some cases, both participants knew all three rules and in other cases there was only partial overlap in their knowledge. Some participants also solved the problem alone to serve as a comparison. The interesting finding was that dyads outperformed individuals only when both participants knew all the rules (i.e., complete overlap). Although there are other potential explanations, it would seem that this is an example of a trade-off between the benefit of diversity and the cost of communication. Simply put, “two heads are better than one” when communication costs do not outweigh the benefits that diversity brings in the form of shared knowledge. Social distribution of cognitive tasks can be beneficial under the right circumstances.

**A design perspective**

One problem common to both ethnographic studies and experiments is that the distribution across persons and artifacts tends to be static. In ethnographic studies of complex cultural practices the distribution of cognition is a result of a long historical development. Although participants might change this distribution in the course of their activity – by inventing new artifacts and processes, for instance – this is quite unlikely. The methods used in experimental studies often prevent participants from altering the distribution of their activity. In experiments, the external media available, what participants are permitted to use them for, and the interaction among participants is usually highly controlled.
Designers are interested in maximising performance in a current practice by changing some of the artifacts, social organisation, or other processes (Kirsh, 1999) and examining the cost-benefit trade-offs that result (Kirsh, 2006). As such, neither ethnographic study of current work practices or strict experiments are particularly useful. The former can only reveal what is done, and the latter will give priority to what could be done without knowing what is currently being done\textsuperscript{41}. As such, Kirsh (2006) proposes doing research using a design perspective. This is somewhat similar in principle to design experiments (A. L. Brown, 1992).

\textit{Epistemic actions.}

To examine how a person and artifact could form a highly reciprocal, coordinated unit Kirsh and Maglio (1994) studied the game Tetris. In this game, players rotate and translate two-dimensional objects (squares, L-shaped, rectangles) in order to fill the space below. The player works on one object at a time that is falling to the bottom of the screen, and this rate of falling increases as the game progresses forcing the player to act more and more quickly. Essentially, the game is about making quick decisions about how to rotate the object and where to place it so that the space below is filled without voids accumulating.

This game has the advantages of being similar to simple tasks like the Tower of Hanoi and can be easily studied and manipulated if needed. As well, it is easy for new players to learn the basic mechanics of the game. In this way, complete novices can be studied as they evolve new practices over a reasonable time period. This is not possible with cultural practices like ship navigation or aircraft piloting. Additionally, there are also existing expert Tetris players who can be compared to these novices. These experts have already developed means of exploiting the game’s affordances to improve their performance.

As Kirsh and Maglio point out, the classical cognitive approach to studying this game would be to assume that players represent the object and the space to be filled below. Then using mental rotation and translation (in working memory) they decide how to act upon the object. And finally, players act on the object using the game controls and

\textsuperscript{41} Additionally, to achieve the desired level of experimental control and manipulation of variables, simple
repeat the previous steps if the action fails to yield the desired outcome (if there’s time). This is not how players play Tetris.

Players rotate objects even before the objects are fully visible (objects drop down from the top of the screen and are revealed in parts). Since players do not know the shape of the object they could not have decided how to rotate the object. Other evidence includes the fact that the players often use rotations and translations that return the object to a previous state. Finally, the actions taken occur far too quickly to have followed mental rotation/translation and decision making. On the basis of these findings, Kirsh and Maglio conclude that often players are rotating and translating objects in order to make the deciding easier, quicker, or less error-prone.

In our view, the failure of classical [approaches] to explain the data of extra rotations is a direct consequence of the assumption that the point of action is always pragmatic: that the only reason to act is for advancement in the physical world. This creates an undesirable separation between action and cognition. If one’s theory of the agent assumes that thinking precedes action, and that, at best, action can lead one to re-evaluate one’s conclusions, then action can never be undertaken in order to alter the way cognition proceeds… To correct this one-sided view, we need to recognize that often the point of an action is to put one in a better position to compute more effectively…

David Kirsh and Paul Maglio (1994, p. 526, emphasis in original)

They label such actions epistemic and contrast them with pragmatic actions. Epistemic actions affect the environment not to move towards the goal directly, which is a pragmatic action, but rather to make the environment more useful for the agent. Specifically, these actions take advantage of the affordances of the environment to make future internal computations by the person more efficient or effective.

The point of taking certain actions, therefore, is not for the effect they have on the environment as much as for the effect they have on the agent.

David Kirsh and Paul Maglio (1994, p. 546, emphasis in original)

In the case of Tetris, using the game to create rotations and translations is simply easier and faster than mental rotation and translation. Kirsh and Maglio also describe other benefits of using these epistemic actions.

The key ideas that Kirsh and Maglio raise (Kirsh, 1999, 2006; Kirsh & Maglio, 1994) are first that there can often be very dense reciprocal interactions between a person and her environment. This interaction goes far beyond the traditional cognitive approach.

“toy” tasks tend to be studied instead of richer, more complex tasks.
of sensing, mentally representing, and acting. Second, that such actions are often taken to affect the person’s future internal computational success.

These epistemic actions could perhaps be thought of as precursors to traditional cultural artifacts that are designed to do the same thing. Whereas epistemic actions arise in the moment of acting and are often ephemeral, cultural artifacts are designed and meant to be durable and persist over months or years. Epistemic actions tend to be individual whereas cultural artifacts are shared. Digital artifacts have interesting properties that could permit epistemic actions of the individual to first become more permanent and then evolve into shared, cultural resources.

**Digital interaction histories**

Hollan and colleagues have proposed that automatic records of the use of digital artifacts (e.g., texts, spreadsheets, webpages, etc.) – what they term ‘histories of interaction’ – could provide useful data for the study of distributed cognition (Hollan et al., 2000, p. 179). This proposal is based in part on earlier work, primarily done by Hill, Hollan and colleagues (Hill & Hollan, 1993; Hill et al., 1992; Wexelblat & Maes, 1999). This work, and the proposed methodology for studying distributed cognition, is discussed more fully below where it is suggested that social software provides a most promising avenue for pursuing this proposal (see Social software and distributed cognition below).

**Complex systems and distributed cognition**

It is interesting to note that although Hutchins (1995a) did not specifically link complex systems with distributed cognition, he modified Simon’s parable of the ant (Simon, 1996, p.51-52) by focusing on a community, or colony, of ants versus a lone ant.
…Simon says, that the trajectory [of the ant on the beach] tell us more about the beach than about the ant. I would like to extend the parable to a beach with a community of ants and a history. … Generations of ants comb the beach. They leave behind them short-lived chemical trails. … Over months, paths to likely food sources develop as they are visited again and again by ants following first the short-lived chemical trails…and later the longer lived roads produced by the history of heavy ant traffic. After months of watching, we decide to follow a particular ant on an outing. This ant seems to work so much more efficiently. … Is this a smart ant? … No, it is the same dumb sort of ant. … But the environment is not the same. It is a cultural environment.

Edwin Hutchins (1995a, p. 169, emphasis added)

This “parable of the ant colony” brings into focus the effects of cultural artifacts in individual cognition, although in the case of ants this is a very special form of cultural artifact; what I will call a stigmergic artifact. Kirsh has also referred to similar phenomena when describing the difference between explicit coordination and implicit coordination.

A second important aspect of coordination is whether it is achieved through explicit - usually symbolic - means, or whether it is achieved through implicit - usually non-symbolic means. …when animals develop trails through the forest, or people develop paths through snow, they are not explicitly coordinating their activity. Trails emerge, the way giant termite hills emerge. Locally optimising behavior leads to global configurations without explicit communication between participants, and without symbolic communication. There are no explicit mechanisms of coordination.

David Kirsh (1999, Coordination section, para. 4, emphasis added)

The above quotes by Hutchins and Kirsh indicate that early proponents of distributed cognition recognised that the interaction of social insects, which is typically described in terms of complex or self-organising systems, is a type of distributed cognition. More recently, some scholars have made direct links between distributed cognition and complex systems (Gureckis & Goldstone, 2006; Heylighen et al., 2004; Poirier & Chicoisne, 2006). For example, Gureckis and Goldstone (2006), who treat groups of individuals as complex systems, have explicitly argued that adaptive group level behaviour should be considered a form of distributed cognition. They argue that what makes something distributed is that it have identifiable units, that are loosely coupled (units are connected, but not totally dependent but not independent either), and the connections among the units are dynamic (Gureckis & Goldstone, 2006). Heylighen (2004) also makes similar arguments for treating distributed cognition in terms of
complex systems. Poirier and Chicoisne connect distributed cognition with emergence, a key characteristic of complex systems: “Truly distributed cognition is emergent cognition,” (Poirier & Chicoisne, 2006, p.217). As examples of distributed cognition they suggest neural networks, ant colonies, and the creation of footpaths by people. The many similarities between complex systems and distributed cognition are summarised in Table 8 below.

Small group collaboration is often clearly a form of distributed cognition (e.g., navigation teams and cockpit crews). However, the parallels with complex systems in such cases are somewhat weak since complex systems typically involve many agents and strong forms of emergence. The similarities between distributed cognition and complex systems are strongest when very large social groups interact using forms of implicit coordination (Kirsh, 1999), or stigmergic interaction.

As previously described, many types of social software derive their benefits from a self-organised process, that is, a large group of people creates useful “knowledge” through stigmergic signals left in a digital environment. The “history-enriched digital objects” research of the early 90’s (Hill & Hollan, 1993; Hill et al., 1992) can now be seen as a form of social software, and is discussed in more detail in the next section.

\[42\] Note that in many simulations that use agent-based models the connectivity in the network is, in fact, not dynamic. Though in real networks that is certainly the case. Social networks are a clear example where the links among the persons involved are constantly changing.
Table 8: Comparing distributed cognition and complex systems

<table>
<thead>
<tr>
<th>Unit of analysis</th>
<th>Distributed Cognition</th>
<th>Complex Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Social groups</strong></td>
<td>Activity distributed across members of a social group</td>
<td>Collection of agents</td>
</tr>
<tr>
<td><strong>Coordination of</strong></td>
<td>Activity distributed across internal and external resources.</td>
<td>Self-organising social systems create stigmergic signs as by-products of individual activity, a form of implicit externalisation</td>
</tr>
<tr>
<td><strong>internal and</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>external resources</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Interaction</strong></td>
<td>Close coupling implies continuous reciprocal interaction of the individual with other individuals and the external resources in the local environment</td>
<td>Feedback loops generate reciprocal interactions and lead to upward and downward causation</td>
</tr>
<tr>
<td><strong>History</strong></td>
<td>Activity distributed over time, or “history matters”</td>
<td>System history is crucial in complex systems since the state of the system at early times affects states at later times</td>
</tr>
<tr>
<td><strong>Emergence</strong></td>
<td>The system’s overall behaviour can not be explained by an understanding of the separate persons and artifacts</td>
<td>Complex systems behave in ways not explained by an understanding of the behaviour of the individual agents</td>
</tr>
</tbody>
</table>
In interaction with objects in the world, history of use is sometimes available to us in ways that inform our interactions with them. For example, a well-worn section of a door handle suggests where to grasp it. A new paperback book opens to the place we last stopped reading. The most recently used pieces of paper occupy the tops of piles on our desk. The physics of the world is such that at times the histories of use are perceptually available to us in ways that support the tasks we are doing. While we can mimic these mechanisms in interface objects, of potentially greater value is exploiting computation to develop new history of interaction mechanisms that dynamically change to reflect the requirements of different tasks.\footnote{James Hollan, Edwin Hutchins, and David Kirsh are currently colleagues in the Department of Cognitive Science, University of California, San Diego.} Digital objects can encode information about their history of use. By recording the interaction events associated with use of digital objects (e.g., reports, forms, source code, manual pages, email, spreadsheets) it becomes possible to display graphical abstractions of the accrued histories as parts of the objects themselves.

James Hollan, Edwin Hutchins, and David Kirsh (2000, p.187, emphasis added)

The metaphor of physical wear as information (e.g., the “history of heavy ant traffic”) (Hill et al., 1992; Hutchins, 1995a), the work on history-enriched digital objects (Hill & Hollan, 1993; Hill et al., 1992; Wexelblat & Maes, 1999), and the recognition of implicit coordination (Kirsh, 1999), appear to have developed in parallel with the distributed cognition framework\footnote{James Hollan, Edwin Hutchins, and David Kirsh are currently colleagues in the Department of Cognitive Science, University of California, San Diego.}.

The work by Hill and Hollan on history-enriched digital objects was based on the metaphor of physical wear (Hill & Hollan, 1993; Hill et al., 1992; Hollan et al., 2000). For example, as a book is used it tends to accumulate wear on important pages. These pages become dog-eared, and the book will tend to open at such pages because of the wear to the book’s spine. The now familiar example of a path through a park is another example of wear. Though the regenerative property of the grass makes this example more dynamic than that of wear in a book, where wear is permanent and irreversible.

Hill and Hollan designed software that would signal to the reader of a document the extent to which other people had read or edited particular lines (Hill et al., 1992). This was indicated using bars in a margin next to the text, for example. Thick bars indicated heavy reading and were based on the amount of time users spent reading that line (estimated from time spent on the overall “page” in a scrolling text). Other forms of digital wear were based on the edits made to the text (e.g., edits to software code). Similar
visual indicators were used to signal to the current user which text section (lines) had been most heavily edited.

Although Hill et al. (1993; 1992) did not clearly state that “history-enriched digital objects” are a form of implicit coordination, this must be the case. Histories, whether the result of physical wear, chemical traces, or digital computation (which is then visually represented), are implicit since the receiving agent generally does not know who the other agents were that caused the history. More importantly, the interactions that generated the marks, wear, chemical accumulation, or “use history” are generally by-products of the previous agents’ activities. Interaction with histories-of-use is, therefore, stigmergic interaction. The artifacts that result are, therefore, stigmergic artifacts.

Many features of social software are derived from maintaining and exploiting digital interaction histories. This form of software clearly produces history-enriched digital objects, and as such, provides an avenue for exploring new forms of distributed cognition that result.

Summary of Distributed Cognition

People can often improve and extend their own cognitive abilities through a coupling with other minds, external artifacts or “cognitive tools” (Lajoie, 2000; Lajoie & Derry, 1993). Distributed cognition provides a framework for thinking about social software that supports self-organised, collaborative text signalling. What Kirsh might describe as a new type of annotation resulting from interaction histories (Kirsh, n.d.). A summary of the literature review follows, which integrates the literature on complex systems, social software, text signalling and distributed cognition.

Summary of the Literature Review

In this review, complex (self-organising) systems were described as systems with multiple agents interacting such that new emergent phenomena occur at the level of the group of agents. Importantly, there is both upward and downward causation such that the agents affect the emerging group level phenomenon and that phenomenon affects the agents. Often the interaction among the agents occurs through stigmergic signs left in the environment, so no direct communication between the agents is required. This permits
asynchronous interaction over different time scales where even the group’s composition may change (i.e., the agents may not be the same).

Also discussed was a new class of online software, called social software, which enables very large groups of people to collaborate. In many cases, this collaboration is indirect and asynchronous. Users explicitly or implicitly attach information to documents or web pages (e.g., tags) that potentially influences other users. These user actions leave stigmergic signs in the information environment. This may lead to a self-organising process whereby the community of users collectively develop metadata. Such metadata is not designed by an author but evolves through the interactions of many learners. Some scholars have begun to explore how social software can be used in education to develop instructionally useful metadata.

The research on text signalling suggests that readers are assisted by the presence of text signals. Text signalling and metadata share much in common. Both are traditionally designed by authors to guide or assist readers and learners. However, most metadata describes entire documents whereas text signals occur within written documents and are local. The local versus global distinction impacts the kinds of tasks the signals or metadata are useful for accomplishing. Global information assists the learner select useful learning resources, for example. Local information may aid the learner’s understanding of the resource by focusing attention and stimulating deeper levels of processing. This difference does not imply that text signals, being local, could not be supported using social software (for local metadata in the form of tags see Bateman, 2007; Bateman et al., 2007; Bateman et al., 2006).

The distributed cognition perspective provides a general way of understanding how many difficult cognitive tasks are successfully completely by a person in combination with her social-cultural environment. This social-cultural environment includes other persons, social conventions, the affordances of the physical environment, and particular artifacts. As such, distributed cognition is a useful framework for investigating how several readers, supported by social software, could produce collaborative text signals that emerge through a self-organised process. In this case, the unit of analysis, when

---

44 Text signals that have a global scope (Lemarie et al., 2008) may be more similar to traditional metadata. For example, an overview occurs at the beginning of a document and describes the entire document at a global level. Metadata, however, are typically shorter and occur outside the document.
investigating an individual reader’s comprehension and performance, must explicitly include the text as a stigmergic artifact.

**Conclusion**

The literature examined has led to the idea that social software could be used to enable readers to generate self-organised collaborative text signals. The readers themselves create the signals through their interaction with the text, augmented with stigmergic signs. Unlike authored text signals, these collaborative text signals are processing “suggestions”, rather than “instructions”, from the community of readers to individual readers. CoREAD, a software application for collaborative reading was designed for this purpose and is described below.
CHAPTER 2: DESIGNING COREAD

Social Software for collaborative reading

CoREAD is a software application (see Figure 5) that was designed and programmed by the author for his doctoral research study (Chiarella & Lajoie, 2006a, 2006b, 2006c, 2007, 2008). CoREAD presents a text in a page-by-page format (no scrolling) and provides the reader with a highlighting function. Font colour, a typographical text signalling device, is used to signal sections of the text depending on the history of highlighting of the group thus far. As such, each reader potentially reads the text with different parts of the text signalled. This is similar to the history-enriched digital objects approach used by Hill and colleagues (Hill & Hollan, 1993; Hill et al., 1992; Hollan et al., 2000). CoREAD uses a weighting function, such that not every reader has the same influence on the text signals. The weighting function will be described more fully below. However, in general, red font indicates sections that have been highlighted by many readers, blue font for sections highlighted by some of the readers, and black font for when few to no readers highlighted the text.
CoREAD was designed to foster a self-organising system of readers where:

1. There are many readers reading a text, making decisions about what sections are important and highlighting them.
2. Readers base their decisions on the text section they are currently reading and do not have any direct interactions with other readers.

Because there are many different forms of mental ability, there are also many different kinds of tests and test items. Some are verbal and others are visual in format. Some tests consist only of abstract-reasoning problems, and others focus on such special competencies as arithmetic, spatial imagery, reading, vocabulary, memory or general knowledge. The broad-spectrum tests, which establish actual IQ scores, typically include a wide variety of items. Before considering these general instruments, however, we must take a brief look at the relations among different specialized tests and how those relations are traditionally interpreted.
3. The text is modified based on the readers’ collective actions. The font colour of the text is altered to reflect the current collective opinion about the importance of each word (i.e., **red** for high importance, **blue** for moderate importance, and **black** for low importance). Technically, for each word an importance score is calculated based on a weighted history (to implement a form of negative feedback, see 6 below) of readers’ actions (i.e., highlighted or not). This score, ranging from 0 to 1, is then converted into one of the text signal states above with the **moderate** category ranging from .30 to <.70.

4. Because the readers have different prior knowledge and reading skills they will each respond to the text differently. This diversity permits varied decisions about the importance of the text sections. In relation to self-organising systems this heterogeneity is a source of randomness, and promotes the modification of the text signals through competing opinions (exploration).

5. By viewing the text signals readers’ decisions about the text may be influenced. If so, this is a form of positive feedback; readers are more likely to attend to and highlight sections currently favoured by the group. This is similar to the way that social insects follow pheromone trails of others that precede them (Bonabeau et al., 1999; Bonabeau & Theraulaz, 2000). This positive feedback loop encourages the use of the text signals by other readers (exploitation).

6. To balance the positive feedback, CoREAD places more weight on the actions of recent readers – a type of negative feedback (but see the Weighting Function section below for more details). This allows any earlier contributions that were not subsequently reinforced by other readers to fade away more quickly (i.e., reset to black font). This negative feedback returns words to a low signal state unless they are consistently highlighted over time and in general prevents the moderate and high text signals from continually growing in number.

   It should be noted that in most engineering situations negative feedback loops are used to achieve desired outcomes (e.g., maintain a state like a temperature setting). Positive feedback loops are to be avoided since they typically cause a cascade that causes the designed system to become uncontrollable [e.g., a single power plant failure that causes
an entire power grid to fail (Ormerod & Colbaugh, 2006). Here positive feedback is beneficial, and is used to generate new, hopefully useful text signals.

If most systems are designed with the control of the individual elements in mind then “designing” a self-organising system is something of a contradiction (Dron, 2007a). A self-organising system in effect designs the emergent phenomenon itself. So what exactly is designed? Here the thing designed is a set of potential interactions among agents and their environment, where the exact outcomes are unpredictable. However, if the types of outcomes that generally occur, or emerge, are predictable then we might well say that something has been designed for a particular purpose. We only need be humble enough to realise that the exact outcome is out of our control. Using positive feedback in CoREAD essentially requires trusting that the readers (the agents), as a group, will bring enough knowledge and skill to the task such that their collective efforts will yield a valuable outcome in most cases.

To recap, self-organising systems as described require 1) many agents, 2) a diversity in agent behaviour or knowledge, 3) readily interpretable information about the agents’ behaviours available in the environment, 4) some degree of positive feedback to allow the system to move toward new states through amplification, and 5) negative feedback to dampen these amplifications and prevent runaway cascading system behaviour. CoREAD should function well in cases where there are many diverse readers (agents) available (i.e., criteria 1 and 2 met). CoREAD provides information about the agents’ behaviours in the text being read which may influence future readers thereby generating positive feedback. CoREAD also employs a negative feedback mechanism.

**Design Choices**

When designing CoREAD the author made several decisions regarding the functions it would support and how information would be displayed to the user. The decisions, and their justifications, are discussed below.

**Typographical text signalling**

Typographical text signals were selected because research demonstrates they typically have positive effects on reading comprehension and recall (Lorch, 1989; Lorch et al., 1995). As well, using the self-organising systems approach described earlier with
typographical text signalling is a tractable problem. The words in the text can be treated as locations on a landscape. This permits the self-organising system properties and functions to be mapped fairly easily to the new context. For instance, “artificial pheromones” (individual highlighting) can be deposited at the word locations. This information, like pheromones, can be allowed to fade with time (the weighting function). The current sum of this information (word importance score) at a word location can then be made visible to the reader in the form of a text signal.

Other forms of signalling (e.g., reader generated headings and previews) could also be employed but would require some additional work, since the mapping would not be as straightforward. This is primarily because these signals would require the construction of new text not merely the selection of existing text, as is the case with highlighting.

**Levels of signalling**

Only two actual signals are used by CoREAD (i.e., high and moderate importance, since the low category is merely the text in its normal state). They are instantiated using one variable, colour. This was done to ensure that the signals were easy for the users to comprehend and remember. It has been found that typographical text signals improve comprehension, memory and search when they are simple and not too numerous (Lorch, 1989; Lorch et al., 1995). For example, studies using complex typographical signals (4 or 5 categories, created by a combination of colour, typeface, and underlining) have failed to show positive effects. In fact, memory performance was worse for both signalled and unsignalled content in these studies (Lorch, 1989, p.224-5).

*Why highlighting?*

The primary reason highlighting was used is because it is location-based in the same way typographical text signals are. Therefore, the former can be used to develop a collaborative history upon which the latter is based. As well, research into students’ spontaneous use of reading strategies shows that highlighting (or underlining, circling, etc) is a natural and common activity among readers with often over 70% using the strategy (Kobayashi, 2007; Lonka, Lindblom-Ylanne, & Maury, 1994; Wade, Trathen, &
Schraw, 1990). By supporting highlighting, CoREAD does not force readers to engage in an unusual activity or one that would unduly affect their normal reading habits.

This use of highlighting was not related to performance in most cases, since it was so commonly used by most students. Some evidence suggests highlighting is beneficial for broad synthesis tasks, however. High school graduates applying to medical school read a 4000 word expository text and were free to take notes and mark the text. Later when asked to explain the title in an essay answer (considered a synthesis task) the more successful students were shown to have used underlining more often (92%) than the less successful students (82%) (Lonka et al., 1994). However, the general finding is that there are few differences between students based on their use of this strategy. Even for the synthesis task discussed, most students, in fact, used the strategy. This leads many to conclude that it is not valuable. Overall, it has been concluded from the research that the use of underlining or highlighting is a poor reading strategy (Pressley & McCormick, 1995, p.480). So why do so many good students use it?

A question to ask is “How effectively are they using it?” Weaker students may highlight irrelevant information and attend to it whereas better students may select important information. More importantly, “What do they do with the highlights while reading a text and afterwards?” Studies often do not permit students to use their notes and highlights after reading during the writing or testing phase (for instance, Lonka et al., 1994). The notes as potential external resources in an extended cognitive system are lost (Clark, 2003). Any benefits the highlights may have afforded the students are lost as well. As a reading comprehension strategy, in cases where students are assessed based on what they can remember about a reading, highlighting may well be a poor choice. However, in authentic contexts where readers know that they will likely reread or refer to longer texts the usefulness of highlighting changes (Shipman et al., 2003). It may not improve understanding on its own, but it is being used as part of a larger strategy. Essentially, highlighting is used to externalise a judgement about what is important. These key passages when so marked make future use of the text more efficient. This is a form of epistemic action (Kirsh, 2006; Kirsh & Maglio, 1994) directed at future work with the

---

45 Smaller studies using richer data have also produced similar results (Marshall, 1998; Marshall & Brush, 2004; Shipman, Price, Marshall, & Golovchinsky, 2003)
text. If this view is correct, then asking students to read a text to answer questions or do a free recall afterwards may not elicit the “evaluate for importance” and then “annotate/mark” activities that the highlighting function ultimately serves. Highlighting a text that you know you will not have access to later may be a pointless exercise. For this reason, a summarisation task was used in the study, following the reading task (see Chapter 4 - Methods below). The summarisation task provided a realistic context for use of the highlighting function that CoREAD provides. Participants were also informed that the text, with their highlights intact, would be available later when the summarisation task took place.

**Weighting Function**

The weighting function that CoREAD uses is not necessarily needed. If no weighting were used text signals would still fade (i.e., font colour reset to black) so long as enough readers eventually chose not to highlight a given section. The weighting function was implemented, however, to minimise the direct contribution that the readers earliest in the history would have on the calculated importance scores. The weighting function, therefore, causes the individual contributions of readers to fade with time. This mimics how chemical pheromones deposited in an environment dissipate over time. It is designed to encourage novelty (i.e., exploration) by giving added weight to recent contributions to the history.

The weighting function adjusts the relative importance of each reader’s contribution to the history. Eventually early readers’ contributions to the history will have almost no direct impact on the importance scores calculated. As more and more readers are added to the history, the early readers’ contributions will only have effects on the outcome (the importance scores and thereby the text signals) through the influence they had on others that was carried forward through positive feedback. For example, the first reader influences the second and third readers fairly directly. These readers can then reinforce the first reader’s actions, thereby extending the first reader’s influence indirectly to the fourth and fifth readers, and so on.

46 The use of overly artificial conditions is a common criticism of reading research. For instance, using short texts with simple structures, and post reading tasks that are impoverished or unfamiliar (Lorch & van den Broek, 1997).
To demonstrate the weighting function, two patterns of highlighting are displayed in Tables 9 and 10 below. For each, the action taken by each reader is shown (H for highlighting, 0 for no highlighting) as well as the importance score that would result under weighting and no weighting (i.e., all readers contribute equally). The pattern in both cases involves initial highlighting by the first one or two readers followed by no highlighting. As can be seen in the tables, the importance scores decrease more rapidly with weighting.

Conversely, if the readers later in the history begin to consistently highlight a word though none of the earlier readers did the weighting causes a higher importance score than no weighting. For example, if the first four readers do not highlight then the next five do, the resulting importance scores are .73 and .56, under weighting and no weighting, respectively. In this case, the later readers have ignored the early readers’ opinions (i.e., “this is not important”) and a high signal emerges under weighting.

Table 9: Effects of the weighting function on the importance scores (text signal state indicated by colour) when only the first of five readers highlights a word

<table>
<thead>
<tr>
<th>Readers</th>
<th>Weighting</th>
<th>No weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>H 0</td>
<td>0.42</td>
<td>0.50</td>
</tr>
<tr>
<td>H 0 0</td>
<td>0.23</td>
<td>0.33</td>
</tr>
<tr>
<td>H 0 0 0</td>
<td>0.14</td>
<td>0.25</td>
</tr>
<tr>
<td>H 0 0 0 0</td>
<td>0.10</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Note: H indicates highlighting, 0 for no highlighting

Table 10: Effects of the weighting function on the importance scores (text signal state indicated by colour) when only the first two of five readers highlights a word

<table>
<thead>
<tr>
<th>Readers</th>
<th>Weighting</th>
<th>No weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td>H H</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>H H 0</td>
<td>0.58</td>
<td>0.67</td>
</tr>
<tr>
<td>H H 0 0</td>
<td>0.38</td>
<td>0.50</td>
</tr>
<tr>
<td>H H 0 0 0</td>
<td>0.26</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Note: H indicates highlighting, 0 for no highlighting
Why use indirect versus direct collaboration?

Direct collaboration, typical in educational settings, requires that groups remain small or the cost of communicating and organising within the groups becomes too high. In small groups, the potential for diversity is limited. Indirect collaboration overcomes such limitations, allowing large numbers of readers to collaborate without increasing communication costs. This allows, though does not guarantee, much more diversity. Diversity is required if a collection of agents is to benefit from interactions in a self-organising system – many identical agents perform no better than a single agent (Page, 2007).

Costs, benefits and trade-offs

In ethnographic studies of distributed cognition (Hutchins, 1995a, 1995b) researchers study a current cultural practice. This practice evolved over time, through explicit design, unintended consequences and the opportunistic use of affordances. One can assume that the resulting practice produces more benefits than costs for the individuals involved. However, when attempting to design changes in a practice – by introducing new social interactions or external tools – the designer needs to consider the costs and benefits of distributing the activity in different ways (Kirsh, 2006). Kirsh proposes several metrics for assessing performance such as speed, accuracy, error type and frequency, learnability and others. As performance improves in one direction it may degrade in another. For example, the cost of communication and organisation when groups of individual collaborate may offset any gains from the increase in diversity of knowledge and skills the group brings to the task. These sorts of trade-offs were quite clearly demonstrated in Zhang’s research (1998) where two individuals only outperformed a single individual when both had the same understanding of the task rules and so communication costs were low.

When designing for an educational context two specific types of trade-offs become important. They are what Pea (1993) describes as a trade-off between 1) access to an activity versus understanding its foundations and 2) a static definition of tasks versus an evolving definition tasks.
Access to an activity

In the first trade-off the individual’s ability to perform without the cultural tool or without the other members of the group is often considered a greater cost than the benefit of being able to participate in an activity. Unfortunately, the typical response to such a trade-off in education is to maintain the status quo, which inevitably means restricting the individual to unaided, “solo”, non-distributed forms of cognition. For example, most assessments must be completed in an unaided fashion or a strictly limited “cheat-sheet” (e.g., a one-page document) is allowed. In some cases access to the activity is considered a greater benefit. An example from education is the use of tables of critical values in statistics. The underlying mathematics used to calculate the critical values need not be taught to students for them to use the values they look up in the tables. These tables, although originally designed to save time, have the advantage of allowing students who might not possess the mathematics knowledge required to calculate the critical values (or more properly the probability of the empirical value they obtained, under the null hypothesis) to perform inferential statistical analyses.

Evolving tasks

The second trade-off concerns viewing the current task in a static fashion without considering that distributing the current task often leads to new and different tasks being within the person’s grasp. That is, by locating an existing cultural tool or by inventing one certain aspects of the current task may be changed. For example, the flight crew may transform a mental computation into a perceptual judgement (Hutchins, 1995b) or a player may use physical rotations to replace time-consuming, error-prone mental rotations in Tetris (Kirsh, 2006; Kirsh & Maglio, 1994). In educational contexts, software can be used to merely perform some tasks more quickly or “neatly” – think of word processors for writing or using graphing software instead of creating hand-crafted graphs. This is the static view of tasks. However, sophisticated graphing software in the hands of someone who understands the fundamentals of graphing can allow for exploratory analyses that could not be easily achieved without it, if at all. For example, creating a three-dimensional graph that can be dynamically rotated for visual analysis cannot be done by hand but requires computer software. Similar outcomes apply to the use of spreadsheets
and word processors. Here the user may take advantage of the affordances of the software to take epistemic actions (Kirsh, 2006) – the use of highlighting to remind the user of edits that need to be done, for example – that make future goal directed (pragmatic) actions simpler or more likely to succeed. These software tools then become more than merely another way of placing marks on paper but rather *alter the process* of writing or calculating or graphing.\(^47\)

Using CoREAD may have unintended costs, and perhaps benefits as well. Reading with CoREAD will potentially affect the access some students have to a text since the collaborative signals may compensate for lack of personal knowledge. Additionally, the affordances CoREAD provides (particularly the visual indicators of importance provided by the collaborative signals) may in fact change the way in which students read a text. There is a possibility that students would use the signals in novel ways. For example, they may choose to skim the text based on these signals.

In any case, the traditional approach of having students read texts “themselves” first before discussion will have to be addressed. As a form of social software, CoREAD challenges some assumptions about education that may make it unappealing to some educators.

**Summary of CoREAD Software**

At a more general level CoREAD can be viewed as an example of an extended mind (Clark, 2001, 2003; Clark & Chalmers, 1998), or more commonly, distributed cognition (Hollan et al., 2000; Hutchins, 1995a; Pea, 1993) in both the socially and artifact mediated sense. The text that CoREAD produces functions as an *artifact* in a distributed cognitive system but is itself also a *product* of the very same distributed cognitive system. The learners are potentially influenced by the text signals presented by CoREAD but also help to modify these same signals: there is both downward (artifact to agent) and upward (agent to artifact) causation. Importantly, readers only interact with each other indirectly through the artifact. It can be said than that when a group of readers uses CoREAD a new text emerges that is a *stigmergic artifact* (text plus signals).

\(^{47}\) Of course this is not to deny that using paper and pen also shapes the process of writing or calculating. However, the ability to program software to respond to changes in the evolving “document” (e.g., change
Using CoREAD is perhaps analogous to passing around a hard-copy document where each reader adds marginalia to the text (annotations and various marks). This results in every reader encountering a slightly different text artifact with its incumbent effects on their processing and comprehension.
CHAPTER 3: THE CURRENT STUDY

If learners in post-secondary institutions are to become successful life-long learners, particularly in professional and learned disciplines, then they require the ability to read disciplinary texts (journal articles, research reports, literary analyses, etc.). Such texts are constantly being published and so the knowledge they collectively contain is therefore very dynamic. As such, typical textbooks can not adequately reflect, in real-time, this changing knowledge base. However, educators have good reasons for using textbooks. Disciplinary texts might prove challenging for many students who may need assistance.

A flexible and adaptive system was therefore sought that would support learners in completing a critical step during the reading comprehension process – determining the relative importance of text segments. Typographical text signals, which emphasise certain text segments, have been shown to improve reading comprehension. However, adding useful text signals to disciplinary texts “by hand” would be very time consuming. When the set of texts is also constantly changing (from year to year, for example) this problem seriously favours the use textbooks – which are static in both content and the signals presented.

There would likely be a great diversity in the relevant conceptual understanding and strategic knowledge a group of students would bring to the task. Could it be possible to combine the knowledge of large numbers of readers, specifically the decisions about importance, in order to automatically add typographical text signals to a text?

The study was designed to investigate the possibility that large groups of students could generate useful typographical text signals through indirect collaboration using a self-organised process. Self-organising systems are often adaptable, flexible, and robust:

1. Adaptable – solutions to problems emerge over time and therefore can adapt to changes in the context of the problem or the available solutions.
2. Flexible – the systems can solve a class of problems (e.g., food foraging) and can respond to a variety of specific situations.
3. Robust – removing or changing some of the individual agents in the system (e.g., adding an agent with poor skills or knowledge) generally does not compromise the performance of the collective.
These features could be useful in educational contexts where static, inflexible texts are more often the norm.

Additionally, since the text signals are produced as a by-product of individual work the efficiency of this approach is very high. The text signals that emerge are free of any costs except the time it takes for a sufficient group of students to read the text. As such, instructors could add new texts in a timely fashion. This would be less likely if a long process involving expert preparation of the texts was required.

**Research Questions**

1. Does indirect collaboration among a group of learners eventually generate stable collaborative signals in a text? How soon does this occur? Is there a pattern to the evolution of the signals?

2. How well do these stable collaborative signals reflect the main ideas in the text?
   2.1. How semantically similar, based on LSA, is an artificial summary (consisting only of the high signals text) to the original text?
   2.2. How does an artificial summary (consisting only of the high signals text) compare to a model summary?

3. Do readers produce better summaries when the text has stable collaborative signals?
   Dividing the readers into pre-stability and post-stability groups:
   3.1. Are the post-stability summaries more semantically, based on LSA, to the original text?
   3.2. How do the post-stability summaries compare to a model summary?

**Hypotheses**

1. The use of CoREAD by a group of heterogeneous readers will eventually lead to stable typographical text signals developed in a self-organised fashion.

2. The collaborative signals indicating highly important text sections (i.e., red font) will tend to correspond with the semantic meaning of the text.

3. The presence of stable collaborative signals in a text will positively influence summarisation. Students from the post-stability period will produce better summaries than those students who read the text before the signals have stabilised.
CHAPTER 4: METHODS

Participants

Ethics approval for the study was obtained from McGill’s Research Ethics Board (see Appendix A) prior to participant recruitment. Participants were recruited by placing ads (see Table 11) on signboards in campus buildings. The sample obtained is therefore a volunteer, convenience sample. Forty (40) undergraduate students participated in the study. They were compensated $40 for participating. The study took between 1.5 and 3 hours to complete.

Table 11: Text from the participant recruitment sign

Recruiting undergraduate students as participants for a research study into reading comprehension.

The main task will involve reading two texts, on topics from psychology, using a novel software tool and writing short summaries for each. The study should take approximately 2 hours.

You will be compensated $40 for your time

The participants were heterogeneous with respect to their degree programs and majors though most had taken a course in psychology or cognitive science at some point (see Table 12). They were generally more experienced undergraduate students with a mean degree year of 3.1 (SD = 1.1). The extent of their relevant prior knowledge with respect to the texts is analysed in detail in the Results section.
Table 12: Participants’ demographic data

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M (SD)</td>
<td>or % as appropriate</td>
</tr>
<tr>
<td>Age</td>
<td>21.9 (2.1)</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>75% female</td>
<td></td>
</tr>
<tr>
<td>Degree Year</td>
<td>3.1 (1.1)</td>
<td></td>
</tr>
<tr>
<td>Degree program</td>
<td></td>
<td>BA- 50%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BSc- 35%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BEd- 10%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BMus- 5%</td>
</tr>
<tr>
<td>Major/Minor in Psychology, a Cognitive Science, or Education</td>
<td>48%</td>
<td></td>
</tr>
<tr>
<td>Number of psychology/cognitive science courses taken</td>
<td>6.9 (9.4)</td>
<td></td>
</tr>
<tr>
<td>Percentage who completed at least one psychology/cognitive science course</td>
<td>73%</td>
<td></td>
</tr>
</tbody>
</table>

These demographic results are based on responses to the Demographic Questionnaires, which are described below in the Demographic Questionnaires subsection of the Methods section.

**Texts**

Two texts were used in the study in order to provide evidence that the results were not text-specific. Descriptions of each text follow below and a summary table is also provided (Table 13).

**Flynn Effect text**

An article from American Scientist (Neisser, 1997) on the Flynn Effect (i.e., the rise in raw IQ scores over time) was adapted for the study (see Appendix B). It is an expository text directed at a broad audience of non-specialists and of a considerable length. It is likely that some of the participants in the study had relevant prior knowledge related to IQ tests. Some may have even been familiar with the Flynn Effect, as it has often been reported in the popular media.

The names of the four sections of the text in the original article were: ‘The ABCs of intelligence’, ‘The Flynn Effect’, ‘Increase or artifact?’, and ‘Explaining the rise’. The first two sections are primarily descriptive, providing information about IQ tests, and describing the Flynn Effect. The last two sections describe hypotheses that explain the Flynn Effect along with evidence for and against. Only two subsections of the fourth section, which described the possible causes of the Flynn Effect, were retained in the version used in the study. In these subsections the author discussed hypotheses relating to
the length of schooling and exposure to visual media, respectively. The author provides arguments supporting the visual media hypothesis and arguments for rejecting the other.

The three introductory paragraphs, headings, and all typographical signals were removed from the text. The actual title was replaced with “FE”. These changes were made in order to remove any text signals originally in the text.

The text as used was 2894 words, 126 sentences in length. The text was presented on 23 “pages” (screens) in CoREAD, each one a paragraph in length, except the first page that simply had the “title”.

Problem Solving text

The fifth chapter of Thinking: An invitation to cognitive science, vol. 3 by Keith Holyoak (1990) on problem solving was adapted for the study (see Appendix C). It is an expository text directed at undergraduate university students and of a considerable length. Sections 5.1.1, 5.1.2, 5.1.3 and 5.1.4 of the chapter were retained in the version used in the study. The headings of these sections were: ‘The nature of problem solving’, ‘Heuristic Search: Means-Ends analysis’, ‘Planning and problem decomposition’, and ‘Production-system models of problem solving’. This was an informational text that simply described problem solving from a cognitive science perspective.

The five introductory paragraphs to the chapter, headings, and all typographical signals were removed from the text. The actual title was replaced with “PS”. These changes were made in order to remove any text signals originally in the text. Slight modifications were made for presentation purposes in CoREAD (e.g., “F^D” was rewritten as “F to the power D”).

The text as used was 2678 words, 128 sentences in length. The text was presented on 21 “pages” (screens) in CoREAD, each one a paragraph in length, except the first page that simply had the “title”. There was a figure in the original text that showed a problem space using a node-link representation (see Appendix D). A paper copy of this figure was provided to the participants when they read the text.
Table 13: Data summarising the characteristics of both texts

<table>
<thead>
<tr>
<th>Text</th>
<th>Original Title</th>
<th>Number of Words</th>
<th>Number of Sentences</th>
<th>Number of Sections used¹</th>
<th>Number of &quot;pages&quot; in CoREAD²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flynn Effect</td>
<td>Rising scores on intelligence tests</td>
<td>2894</td>
<td>126</td>
<td>4</td>
<td>23</td>
</tr>
<tr>
<td>Problem Solving</td>
<td>Problem Solving</td>
<td>2678</td>
<td>128</td>
<td>4</td>
<td>21</td>
</tr>
</tbody>
</table>

¹ Sections were not separated or visually indicated in CoREAD and section headings were removed.
² Each page or screen in CoREAD displays an entire single paragraph (no scrolling is required), except for the first screen which displayed a “title” (simply FE and PS, respectively).

Software

CoREAD, as previously described, was used to display the texts. The texts appeared one paragraph at a time onscreen with no scrolling required (entire paragraphs were visible on each “page”). Collaborative text signals were displayed during reading but removed during the summarisation stage (see Procedure below). Users had complete freedom to move forward or backward one page at a time or could also go directly to any particular page of the text using a slider.

Demographic Questionnaires

The demographic questionnaires, one for each text, were designed to 1) collect basic demographic data, 2) determine if the participants had completed coursework related to the texts, and 3) determine if the participants were already familiar with the specific topics of the texts (see Appendices E and F). To assess their prior knowledge, the participants were asked to indicate on a six-point scale (1 = “very low to none”, 6 = “very high”) their level of knowledge about some of the main concepts in the texts. These concepts included ‘the Flynn Effect’ and ‘Possible causes of the Flynn Effect’, and ‘Problem solving as search’ and the ‘Problem space metaphor (states, operators, paths)’ which were the main ideas from the Flynn Effect and Problem Solving texts, respectively. The questionnaires were completed after the participant had both read the text and written

⁴⁸ Please note that this is a self-report indicating the participant’s perceived level of knowledge, and should not be treated as a proper assessment of such knowledge.
the summary. This order was used so that the questionnaire would not influence how the participants read the text or wrote the summary.

**Equipment Set-up**

The participants read each text and wrote the respective summaries at a large desk. A Toshiba M200 tablet PC was placed on the desk on an inclined stand to provide a more comfortable writing angle. The tablet PC is controlled using a stylus, a pen-like device that also functions like a mouse. Participants read the texts using this tablet PC. When reading the text the participants could highlight sections by simply touching the stylus to the screen and dragging it across the screen. Highlighting was therefore executed in a fairly natural manner using the stylus. Unlike a real highlighter, however, sections of text spanning several lines could be highlighted by dragging the stylus from the beginning of the section directly to the end (like using a mouse). This is much easier than highlighting each line separately. Additionally, since CoREAD highlights whole words, participants could highlight whole words by simply tapping them with the stylus.

An external LCD monitor was also connected to the tablet PC. This allowed the researcher to view each participant’s progress through the texts. An external keyboard and mouse were also connected to the tablet PC and located in front of the external monitor. The researcher used these to control the tablet PC as needed.

A standard laptop, with a mouse, was provided to the participants for writing the summary. These were placed on the desk beside the tablet PC during the summarisation phase (see Procedure below). The participants could operate both the tablet PC and the laptop simultaneously allowing them to consult any page in the text while writing the summary.

**Procedure**

Participants were instructed to read and highlight the text in order to write a summary. They were also informed that the text and their highlights would be available during the writing phase. This was done to ensure that they carefully read the text and used highlighting in a manner consistent with and useful for writing a summary. Each participant completed the study separately in sequence. Therefore, participant one is also the first participant in the collective history, the last participant the 40th.
The order of the texts was counterbalanced; odd numbered participants read the *Flynn Effect* text followed by the *Problem Solving* text, the reverse order was used for even numbered participants. This was done to ensure that the order of presentation did not consistently affect one text. For example, participants might read the second text more quickly or easily after having used CoREAD once before. As well, the participants might experience fatigue since the study was quite long and demanded their full attention. This would also affect how the second text was read or summarised.

For each text the following procedure was used (except for steps one and two, which were completed only once):

1. Participants were informed of their rights and then read and signed the Consent Form (Appendix G).
2. Participants viewed a video demonstration of how to use CoREAD.
3. The text was displayed on the tablet PC and the reading instructions were read aloud to the participant (see Table 14). These task instructions, printed on a card, were also given to the participant for later reference, if needed (see Table 15). For the *Problem Solving* text, a figure of a problem space that was printed on a card was handed to the participant as well (see Appendix D).

Table 14: Verbal instructions: Reading Task

```
As you can see, [number] students have read this text before you. The font colour of the text - red, blue, or black - is therefore based on the highlighting of these [number] students. Red indicates... Blue indicates… Black indicates…

Please read the entire text keeping in mind that you will write a short summary of the text when you are done. You will have access to the text, with your highlighting intact, while you write the summary. However, the text will be displayed in plain black - the importance information will **NOT** be visible while you write the summary.

Here are the task instructions <hand card> as a reminder.”
```
Table 15: Reading Task Instruction card

- Please read the entire text keeping in mind that you will write a short summary of the text when you are done.
- You will have access to the text, with your highlighting intact, while you write the summary.
- However, the text will be displayed in plain black - the importance information will not be visible while you write the summary.

4. The participant read the entire text using the highlighting function as she wished. There was no time limit for the reading task.

5. The participant was instructed as to the nature of the writing task (see Table 16) and provided a card with the instructions for later reference as needed (Table 17). The participant wrote a summary of the text on the laptop. As such, the participant could not copy and paste from the text to her summary. The word processor file was designed so that the summary could be completed on one page and this page was entirely visible onscreen without scrolling. The page allowed a maximum of approximately 420 words to be written. The text to be summarised – with the participant’s highlighting intact but with the collaborative signals removed – was available to the participant while writing the summary.

Table 16: Verbal instructions: Writing Task

"Please write a summary of the text in this WORD file. You may consult the text on the Tablet PC as you write the summary. The summary may be a maximum of 1 page in length.

Here are the task instructions <hand card> as a reminder."

Table 17: Writing Task instruction card

- Please write a summary of the text in this WORD file.
- The summary may be a maximum of 1 page in length.
- You may consult the text on the Tablet PC as you write the summary.

6. The participant completed the Demographic Questionnaire (see Appendices E and F) appropriate for the text just summarised.

The researcher remained in the room with the participant for the duration of the study and monitored the participant’s progress. Both texts were read in a single session lasting about 1.5 – 3 hours. Participants were debriefed at the end of the study and then compensated.
**Data Sources**

Various primary data sources were collected as listed below. The first two data sources were collected by CoREAD. The researcher recorded the start and end times for the writing tasks.

1. a record of the entire history of 1) highlighting, 2) the importance scores, and 3) the text signal states for all words across all participants
2. the sections of the text highlighted by each participant
3. the total reading time
4. the written summary
5. time spent writing the summary
6. the demographic questionnaire data
CHAPTER 5: RESULTS

Basic results characterising the study

Summary statistics for both texts are displayed in Table 18 below. These results indicate that most participants spent about 50 minutes with each text. This suggests that an adequate amount of time was spent comprehending the texts and writing the summaries. Also, the summaries were, in general, sufficiently long to cover the main ideas in the texts. Finally, there were no differences between the texts in the amount of time spent or the length of the summaries written.

Table 18: Summary statistics (M and SD) for both texts [n=40]

<table>
<thead>
<tr>
<th></th>
<th>Flynn Effect text</th>
<th>Problem Solving text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading Time (minutes)</td>
<td>23.8 (7.5)</td>
<td>23.0 (7.1)</td>
</tr>
<tr>
<td>Writing Time (minutes)</td>
<td>28.1 (10.2)</td>
<td>27.6 (9.4)</td>
</tr>
<tr>
<td>Length of Summary (words)</td>
<td>372.9 (65.7)</td>
<td>373.6 (55.3)</td>
</tr>
</tbody>
</table>

* Due to data loss for one participant, n=39

Participants’ Prior Knowledge

The responses to the Demographic Questionnaires (Appendices E and F) about the courses taken and prior knowledge items were analysed; the results by text are presented below. Recall that these are based on the participants’ self-assessments of their own knowledge of the concepts and not the results of a test.

Flynn Effect text

Many of the participants had studied statistics (70%) and some had studied intelligence as a major topic in a course (35%) (see Figure 6 below). There was quite a lot of variability in their knowledge of statistics concepts (e.g., ‘correlation’) and IQ concepts (e.g., ‘types of intelligence tests’) (see Figure 7 below). They also reported having little knowledge of ‘the Flynn Effect’ or ‘Causes of the Flynn Effect’. Only two participants reported having “high” or “very high” knowledge of the ‘the Flynn Effect’. These results indicate that the sample was quite heterogeneous with respect to their relevant prior knowledge, though most of the participants had not learned of the Flynn Effect prior to the study.
Figure 6: Course/Major topics studied related to the Flynn Effect text
Figure 7: Prior knowledge related to the Flynn Effect text
Some of the participants had studied cognitive psychology (45%) but few had studied problem solving (20%) or artificial intelligence (15%) as a major topic in a course (see Figure 8 below). For the most part they reported having little knowledge of the concepts related to the Problem Solving text (see Figure 9 below). For most concepts the most common response was “very low to none”. Items for which there was more diversity of knowledge included ‘Problem solving as search’, ‘Problem decomposition’, and ‘Exhaustive search’. These results indicate that the participants had less relevant prior knowledge related to the Problem Solving text than they did for the Flynn Effect text. As such, one might expect this to have been the more difficult text.

Figure 8: Course/Major topics studied related to the Problem Solving text
Figure 9: Prior knowledge related to the Problem Solving text
Summary of Participants’ Prior Knowledge results

The participants reported having more prior knowledge related to the Flynn Effect text than they did for the Problem Solving text. Recall, however, that participants spent the same amount of time reading and writing the summaries for both texts. Very few participants reported “high” or “very high” knowledge for ‘the Flynn Effect’ and ‘Possible causes of the Flynn Effect’, and ‘Problem solving as search’ and ‘The problem space metaphor’ which indicates that for most participants these texts covered new content.

General outcomes

Before addressing the research questions, outcomes that could undermine the results are examined. These concern how CoREAD was used by the participants and the general nature of the collaborative text signalling outcomes. The results below indicate that none of these outcomes confound or negatively affect the interpretation of the main findings.

Did participants use the highlighting function?

On average participants highlighted 29% ($SD = 10\%$) of the Flynn Effect text and 28% ($SD = 12\%$) of the Problem Solving text (also see Figure 10). The use of this feature varied quite a lot as the standard deviations indicate. Some participants highlighted very little (e.g., participants 21 and 35 for the Problem Solving text) while others highlighted nearly half the text (e.g., participants 4 and 17 for the Problem Solving text).

Did the participants’ use of highlighting grow over time?

Though the amount of highlighting varied across participants it had no discernible pattern over time (see Figure 10). This is important because it indicates that the participants did not change their propensity to highlight the text in response to the presence of the text signals.

Did the amount of text signalling increase over time?

Any concern that the percentage of the text signalled might grow with time is unfounded. This could be a concern because previous work has indicated that typographical text signals are ineffective if they are too numerous (Lorch et al., 1995). As
the graphs indicate (Figure 10) the percentage of moderate signals in the text levels off quite early on. High signals actually reduced somewhat over time beginning around the 20th participant. The mean percentage of the text in the three states is presented in Table 19 below. The negative feedback mechanism in CoREAD seems to regulate the amount of signals in the text, preventing them from growing uncontrollably. The percentage of text in the high signal state is still quite high compared to the Lorch study (1995), however. Future research will be needed to ascertain if the parameters in CoREAD need to be changed so as to lower the amount of these signals so as to improve their usefulness.

Table 19: Mean (SD) percentage of the text displayed to the participants by signal state

<table>
<thead>
<tr>
<th>Signal state</th>
<th>Flynn Effect text</th>
<th>Problem Solving text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>60% (5%)</td>
<td>59% (5%)</td>
</tr>
<tr>
<td>Moderate</td>
<td>24% (2%)</td>
<td>28% (3%)</td>
</tr>
<tr>
<td>High</td>
<td>17% (3%)</td>
<td>14% (4%)</td>
</tr>
</tbody>
</table>

Note: statistics are based on participants 2-40 for high signals and participants 3-40 for moderate signals
Figure 10: Percentage of the text in each signal state, as viewed by the participant, and the amount of text highlighted by the participant.
Were participants overly influenced by the collaborative text signals?

Agreement was consistent with the participants being influenced by the text signals though a proper controlled study will need to be performed to establish actual influence: it is possible that these levels of agreement were due to independent actions. As would be expected, participants highlighted the high signal text more than the moderate or low signal text (see Table 20). Although the mean percentage highlighted for the high signal text is quite high (77% and 73%) it still indicates that the participants did not completely agree with the collaborative text signals; they chose not to highlight (i.e., disagreed) in 23% and 27% of the cases. Similarly, for the low signal text participants only highlighted on average 9% and 10%. This also indicates high agreement with the collaborative text signals (i.e., by not highlighting low signal text most of the time) but also that there were times when they disagreed.

Table 20: Mean (SD) percentage of text highlighted by signal state

<table>
<thead>
<tr>
<th>Signal state</th>
<th>Flynn Effect text</th>
<th>Problem Solving text</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>77% (19%)</td>
<td>73% (23%)</td>
</tr>
<tr>
<td>Moderate</td>
<td>47% (21%)</td>
<td>47% (23%)</td>
</tr>
<tr>
<td>Low</td>
<td>9% (7%)</td>
<td>10% (8%)</td>
</tr>
</tbody>
</table>

Note: statistics are based on participants 2-40 for high and low signals and participants 3-40 for moderate signals.

From another viewpoint, Figure 11 below shows that most participants highlighted a combination of high, moderate, and low signal text and that this did not change over time. This indicates that new, individual contributions to the collective history were being made throughout the study. Any concern that an information cascade (Surowiecki, 2004) would occur, whereby participants later in the sequence would no longer contribute their own opinions but rather defer to the group, is unwarranted.
Figure 11: Total number of words highlighted by each participant coded by signal state
Correlational analyses of basic measures

To examine the relationships between some of the basic measures Pearson’s correlations were computed for all pairs of the following variables:

1. total reading time
2. prior knowledge score
3. time to write summary
4. length of summary (words)
5. LSA cosine
6. LSA vector length

The prior knowledge items were assigned scores from 1 to 6 by category of response: 1 for “low to none” and 6 for “very high”. The prior knowledge score was then calculated by summing the scores across all items. The maximum score for the Flynn Effect text was 66. For the Problem Solving text the maximum was 60.

The results of the correlation analyses for both texts are displayed in Tables 21 and 22 below. For both texts reading time was positively related to writing time, perhaps indicating a difference between participants regarding their commitment to the task or motivation to perform well. Or it may mean students who had difficulty with the texts took more time on both tasks.

As would be expected, the length of the summaries for both texts was strongly and positively related to LSA vector length and LSA cosine. The vector length indicates the amount of semantic content contained in the summary and it is well known to be correlated with the number of words. LSA vector lengths were in turn positively related to LSA cosine.

For the Flynn Effect text the writing time was also positively related to LSA cosine. In the case of the Problem Solving text, the prior knowledge score was negatively related to reading time. However, removing an outlier (reading time 47 minutes, approximately twice the mean) resulted in a non-significant correlation.

49 The LSA cosine and LSA vector lengths resulted from the analyses of the participants’ summaries compared to the texts (see Results – Latent Semantic Analyses and Research Question 2 below for details).
Table 21: Statistically significant* correlations between basic measures for the Flynn Effect text

<table>
<thead>
<tr>
<th></th>
<th>Reading time</th>
<th>Prior knowledge</th>
<th>Writing time</th>
<th>Summary length</th>
<th>LSA cosine</th>
<th>LSA vector length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior knowledge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Writing time</td>
<td>+ 0.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summary length</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSA cosine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSA vector length</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* all reported correlations are statistically significant at p<.05

Table 22: Statistically significant* correlations between basic measures for the Problem Solving text

<table>
<thead>
<tr>
<th></th>
<th>Reading time</th>
<th>Prior knowledge</th>
<th>Writing time</th>
<th>Summary length</th>
<th>LSA cosine</th>
<th>LSA vector length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior knowledge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Writing time</td>
<td>- 0.34</td>
<td></td>
<td></td>
<td></td>
<td>+ 0.60</td>
<td></td>
</tr>
<tr>
<td>Summary length</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSA cosine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+ 0.69</td>
<td>+ 0.91</td>
</tr>
<tr>
<td>LSA vector length</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+ 0.76</td>
<td></td>
</tr>
</tbody>
</table>

* all reported correlations are statistically significant at p<.05

Summary of general outcomes

These results show that the participants used the highlighting function and that this use did not grow over the time. The presence of the text signals did not encourage participants later in the sequence to highlight more and more of the text. Similarly, the
text signals themselves, particularly the high signals, did not increase over time either. This is important since too much signalling would lower the usefulness of the individual signals (Lorch et al., 1995). Last, the participants seemed to agree with each other, suggesting an influencing effect of the text signals. However, there was disagreement as well, indicating that participants continued to use their own judgement and added new “information” to the collective history. There appeared to be a balance of exploitation and exploration. In sum, CoREAD was used in a manner and produced text signalling outcomes that should not pose a concern to the results that follow.

**Research Question 1 – Emergence of stable typographical text signals**

This section addresses the first research question;

Does indirect collaboration among a group of learners eventually generate stable collaborative signals in a text? How soon does this occur? Is there a pattern to the evolution of the signals?

Deciding when a pattern has emerged, which is relatively stable, is not straightforward. Indeed, in some studies the researchers present snapshots of the changing pattern represented visually and rely on the reader’s ability to perceive patterns rather directly (for examples see Goldstone & Ashpople, 2004; Goldstone & Janssen, 2005; O. Holland & Melhuish, 1999; Schelling, 2006). Metrics that summarise the state of the system, calculated for each time period\(^{50}\), are also often used.

To answer the first set of research questions the data were analysed in different ways. Stability is the lack of change, and occurs when a pattern holds constant for a certain period of time. As such, the history of the signals was examined using three methods:

1. The percentage of the words that changed signal state (e.g., from high to moderate) from one participant to the next.
2. The percentage of the words remaining in the same signal state over a fixed number of participants.
3. The changes in the agreement levels (using Percent agreement, Cramer’s V, and Pearson’s correlation) between points in the history, where the level of

\(^{50}\) These time periods may correspond to rounds if the participants (or agents in a simulation) act in a specified sequence or they may simply be set time intervals (e.g., 5 minutes)
agreement indicates similarity in the state of the signals at two times.
Increasing stability would result in agreement levels between time points of a
fixed distance increasing. In other words, there should be more agreement
between the signals for the 40th and 30th participants than for the 30th and 20th,
if say, the signals stopped changing much after the 30th participant.

**Percentage of the words that changed signal state (Method 1)**

To examine stability, the percentage of changes in the text signals *that occurred
after each participant* was examined. If a word changed from one signal state to another
this was coded as a change (e.g., from blue to red, moderate to high importance). Graphs
of these changes, one for each text, are presented in Figure 12 below. These graphs
indicate that the changes generally decrease over time, which is indicative of increasing
stability. Power functions seem to fit these data fairly well ($R^2=.88$ and .86 for the Flynn
Effect and Problem Solving texts, respectively). The changes decrease fairly rapidly. After
participant 30 the percent changes range between 0.7% -1.5% for the Flynn Effect text
and 0.7% - 2.1% for the Problem Solving text.

**Percentage of the words in the same signal state over eight participants (Method 2)**

To examine stability, the percentage of words that were in the same signal state
over the last eight (8) participants was assessed. Eight was used since it is equivalent to
20% of the total sample. If a word changed from one signal state to another at anytime in
the last eight participants (including the current participant, e.g., 1 to 8, or 33 to 40) this
was coded as unstable. Those words that remained in the same state were coded for that
state (i.e., high, moderate, or low). This is a much stricter criteria for stability than the
previous analysis. Graphs for both texts are presented in Figure 13 below. The graphs
show the percentage of the text that was stable over the last eight participants by state
(low, moderate and high) and the percentage that was unstable. These results suggest that
after the 30th participant, for both texts, the high signals remain quite stable. The unstable
portions of the text are also quite low by this time and decrease again over the last few
participants.
Figure 12: Percent change in the collaborative text signals after each participant
Figure 13: Percentage of the collaborative text signals that remained the same over the last eight participants by state (low, moderate, high) and those that changed (unstable)
Changes in the agreement between points in the history (Method 3)

The changes in the agreement (using Percent agreement, Cramer’s V, and Pearson’s correlation) between points in the history were computed for distances in the collaborative history of 10, 20 and 30. Each measure of level of agreement indicates the similarity in the text signals at two times. A pattern of increasing stability would result in agreement levels, between time points of a fixed distance, increasing as we compare participants later in the history. In other words, there should be more agreement between the signals for the 40th and 30th participants than for the 30th and 20th, if the signals get increasingly stable over time.

Percent agreement and Cramer’s V were calculated by comparing the actual signal states (low, moderate, and high) across all words at two points in the history. Cramer’s V is derived from Chi-square which was computed first. The formula for Cramer’s V is

$$Cramer’s\ V = \frac{X^2}{N(L-1)}$$

where N is the sample size (in this case the number of words) and L is the smallest dimension of the table (in this case there were 3 rows by 3 columns, therefore L = 3). Cramer’s V ranges between 0 and 1, where 1 is perfect association.

Pearson’s correlations were calculated by using the importance scores of the words at the same points in the history. Note that these scores contain more information than the signal states.

The author is aware that the assumption of independence is certainly violated in these analyses. However, the Chi-squares and correlation coefficients are not being used to reject null hypotheses and so this assumption is irrelevant here. The value of the statistics, namely Cramer’s V (derived from Chi-square) and the correlation coefficients will be compared over “time” in the collaborative history to see if any patterns hold.

The results for both texts are presented below in Tables 23 and 24, respectively. Comparing the measures of association from left to right one will see that they increase (see the top two rows, the third row has just a single cell). This increase occurs in spite of the fact that the distance is kept constant. This increase also occurs for both texts. The pattern of increasing association as one compares later points in the history indicates increasing stability of the text signals. For the Flynn Effect text, for instance, there is a
96% match (percent agreement) between states of the text signals when comparing participant 30 and 40. However, comparing participants 10 and 20 only produces an 89% match (percent agreement). As would be expected, the greatest distance (participant 10 versus 40) produces the lowest scores in general in both texts.

Table 23: Three measures of association (Percent agreement, Cramer’s V, Pearson’s correlation) for equidistant points in the history (by table row) – *Flynn Effect* text

<table>
<thead>
<tr>
<th>Distance</th>
<th>10-20</th>
<th>20-30</th>
<th>30-40</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>89%</td>
<td>91%</td>
<td>96%</td>
</tr>
<tr>
<td>10-20</td>
<td>0.79</td>
<td>0.83</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>0.94</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>10-30</td>
<td>84%</td>
<td>20-40</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.70</td>
<td>0.81</td>
<td>0.96</td>
</tr>
<tr>
<td>10-40</td>
<td>84%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.70</td>
<td>0.90</td>
<td></td>
</tr>
</tbody>
</table>

Note: the cells are for comparisons between participant 10 versus 20, 20 versus 30 etc.

Table 24: Three measures of association (Percent agreement, Cramer’s V, Pearson’s correlation) for equidistant points in the history (by table row) – *Problem Solving* text

<table>
<thead>
<tr>
<th>Distance</th>
<th>10-20</th>
<th>20-30</th>
<th>30-40</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>85%</td>
<td>91%</td>
<td>95%</td>
</tr>
<tr>
<td>10-20</td>
<td>0.72</td>
<td>0.83</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>0.91</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>10-30</td>
<td>82%</td>
<td>20-40</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.67</td>
<td>0.80</td>
<td>0.95</td>
</tr>
<tr>
<td>10-40</td>
<td>83%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.68</td>
<td>0.87</td>
<td></td>
</tr>
</tbody>
</table>

Note: the cells are for comparisons between participant 10 versus 20, 20 versus 30 etc.
**Results Summary: Research Question 1**

Three methods were used to examine stability. In all three cases it was demonstrated that CoREAD seems capable of generating stable typographical text signals using indirect collaboration. In addition, the signals seem to reach a fairly high level of stability by soon after the 30th participant for both texts. This is interesting since earlier analyses suggested that the *Problem Solving* text might have been more difficult for the participants. It might have been expected that the signals for this text would have stabilised more slowly, this does not appear to be the case. In summary, readers using CoREAD form a self-organising system that generates stable, emergent typographical text signals.

**Latent Semantic Analysis**

Latent Semantic Analysis (LSA) was used to compute text similarity to address research questions 2 and 3 (Foltz, 1996; Foltz, Kintsch, & Landauer, 1998; Landauer, 2002; Landauer, Foltz, & Laham, 1998; Landauer, McNamara, Dennis, & Kintsch, 2007). A review of the technique and some relevant empirical findings follow.

*Creating a latent semantic “space”*

Although LSA can be viewed as a model of semantic representation, a theory of meaning (Landauer, 2002; Landauer et al., 2007), it can also be treated as a practical method for analysing text meanings (Foltz et al., 1998). LSA was used as a practical method in this study to assess the semantic similarity of various texts. For example, to compare the collaborative signals with the text itself and the participants’ summaries of the text with the text itself.

Latent semantic analysis requires a large text corpus. This corpus is then divided into passages, usually paragraphs. A word by passage matrix is then constructed where each cell is the frequency with which a word appeared in the passage. This matrix is inherently sparse, having mostly 0’s. Cell frequencies are weighted (log entropy weighting) such that words that are very frequent across passages (e.g., prepositions, articles, etc.) are adjusted so that they are less important, since they carry little semantic content. This matrix is then decomposed into three matrices: a term (word) matrix, a
diagonal matrix, and a passage matrix. When these three matrices are multiplied together
the original matrix is reconstructed. At this point, the dimension of the solution is usually
reduced by eliminating (setting to 0) the lowest values in the diagonal matrix (these
values are ordered from highest to lowest along the diagonal, all other values are 0). This
step is similar to choosing to keep the initial components in a principal components
analysis. Because of the reduction in dimensionality LSA can infer that two words are
related even if those words never appeared together in any passage in the corpus. The
reduction step in some sense removes the noise in the data leaving only the most
important dimensions, those that separate the words and passages based on their
underlying meaning. Unlike principal components analysis, the number of dimensions
(factors) retained in LSA is usually several hundred. Landauer et al. (1998) provide a
simple example of LSA that is described next. In this simple case, the semantic space
consisted of nine titles about two different topics (9 passages) with 12 words\(^{51}\) (9 x 12
matrix). Retaining only two dimensions LSA correctly inferred the two separate topics
producing:

- high positive correlations\(^{52}\) among the 4 titles of the first topic
- high positive correlations among the 5 titles of the second topic, and
- high negative correlations between titles from different topics

Once a semantic space has been generated it can be used to compute word-word, word-
passage, or passage-passage similarity.

**Using LSA to compute similarity**

LSA can compute the semantic similarity between two pieces of text given a
semantic space. For each text, the words it contains are added together; specifically, their
vectors are added together. Therefore, each text is represented by a vector that is equal to
the sum of vectors of the words it contains. Similarity is then calculated by computing the
cosine for the two vectors representing the two texts. This cosine metric may range
between -1 and +1, where +1 is maximum semantic similarity (similar to a correlation
coefficient but the cosine has a different mathematical basis). In physical space, two

\(^{51}\) Note that only words that appeared in a least two titles were retained for the initial matrix.

\(^{52}\) Spearman correlations were used.
vectors (of equal length and common origin) that meet at the following angles have the indicated cosine:

- $90^\circ = \cos 0$, (perpendicular)
- $180^\circ = \cos -1$ (directly opposite)
- $0^\circ = \cos 1$ (perfect overlap)

LSA is sometimes described as a “bag of words” approach, as it does not use word order, syntax, or even sentence boundaries in the analysis. It does have certain limitations. For instance, two texts, sentences for example, that are opposite in meaning because of a negation such as “not” would be very similar (very high positive cosine) when compared using LSA. The negation carries little semantic content on its own and so it contributes little to the LSA calculation.

Overall, LSA has proven capable of providing valid measures of semantic similarity (Landauer et al., 1998; Landauer et al., 2007). One such use of LSA is the automated grading of essays.

*Scoring essays with LSA*

LSA has been used to score essays and has been incorporated into a tool specifically for that purpose called the Intelligent Essay Assessor\(^53\) (Foltz, Laham, & Landauer, 1999). Studies have indicated that the reliability of LSA to score essays is as good as human inter-rater reliability. Essay scoring using LSA is performed by comparing the essays with 1) other essays previously scored by human raters, 2) a model essay written by an instructor or similar, or 3) the original text(s) about which the essays were written (Foltz, 1996; Foltz et al., 1999).

*Why use LSA?*

LSA was used in this study for several reasons. The texts used are very long, much longer than those typically used in reading research, and too long for a fine-grained propositional analysis. LSA has been shown to be as reliable as human judges when rating essays – the inter-rater reliability with human judges is as good as between the judges themselves. LSA provides an objective measure that is reproducible by others.

\(^{53}\) The Intelligent Essay Assessor is commercial software available from Pearson Education.
(replication), and any future studies with CoREAD using these texts could be directly compared to the results that follow.

Were the LSA spaces used adequate?

Before presenting the results for research questions 2 and 3 the adequacy of the semantic spaces used for the latent semantic analyses is assessed. This was examined using the ‘Near Neighbors’ application. This application returns two lists of words. First, words from the semantic space that were most similar to the text submitted (in this case the texts used in the study). Second, words that appeared in the text submitted but were missing in the corpus that was used to construct the semantic space. Word types that do not exist in the corpus are ignored in the texts submitted for analysis and therefore make no contribution to the meaning of the text submitted.

Analyses were performed using the LSA software provided by the SALSA group at the University of Colorado at Boulder (see Landauer et al., 2007). The Flynn Effect text analyses were performed using the psychology semantic space (Myers, 1995) using all 400 factors. The Problem Solving text analyses were performed using the cognition semantic space (labelled ‘cognit’) using all 300 factors.

Flynn Effect text

The psychology semantic space is based on Myer’s psychology textbook (1995), which includes a chapter on intelligence. The textbook corpus included 4903 paragraphs by 19160 unique terms (reported in Foltz et al., 1998). The Flynn Effect text was submitted to a ‘Near Neighbours’ analysis. The top 10 terms most semantically related to the Flynn Effect text are shown in Table 25 with their cosine measure. These terms include the names of two common IQ tests, the words “IQ” and “score(s)”, and the name “Flynn”. This indicates that the Myers textbook covered the topic in question.

---

54 Note that in the Foltz et al. study only 300 factors were retained.
Table 25: A ‘Near Neighbours Analysis’*: Top 10 terms most similar to the Flynn Effect text

<table>
<thead>
<tr>
<th>Term</th>
<th>Similarity (cosine)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Wechsler</td>
<td>0.50</td>
</tr>
<tr>
<td>2. WAIS</td>
<td>0.48</td>
</tr>
<tr>
<td>3. Score</td>
<td>0.47</td>
</tr>
<tr>
<td>4. Flynn</td>
<td>0.42</td>
</tr>
<tr>
<td>5. Standardization</td>
<td>0.41</td>
</tr>
<tr>
<td>6. Scores</td>
<td>0.40</td>
</tr>
<tr>
<td>7. IQ</td>
<td>0.40</td>
</tr>
<tr>
<td>8. Counters</td>
<td>0.40</td>
</tr>
<tr>
<td>9. Changeable</td>
<td>0.40</td>
</tr>
<tr>
<td>10. Restandardized</td>
<td>0.40</td>
</tr>
</tbody>
</table>

* No terms were removed and pseudodoc (log entropy weighting) was used.

The following “missing words” list was returned for the Flynn Effect text. The only missing words of some concern are “raven” and “matrices”.

List of missing words: accounted, albeit, artifact, attains, attendance, cahan, christened, cliche, competencies, consisted, counties, decorations, diagrams, dolts, ensures, extrapolating, grandfathers, greenfield, indexed, jerusalem, matrices, midst, montages, normalized, otago, paradox, preposterous, psychometricians, raven, renaissance, secular, sheer, sheets, shining, sophistication, sorel, static, successors, superseded, tester, undiminished, unitary, universality, wrappers

Problem Solving text

The Problem Solving text was submitted to a ‘Near Neighbours’ analysis. The top 10 terms most semantically related to the Problem Solving text are shown in Table 26 with their cosine measure. These terms include “goal”, “solution” and “subgoal” and the words “problem” and “solving”. This indicates that the cognition space covered the topic in question reasonably well.
Table 26: A ‘Near Neighbours Analysis’: Top 10 terms most similar to the Problem Solving text

<table>
<thead>
<tr>
<th>Term</th>
<th>Similarity (cosine)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal</td>
<td>0.44</td>
</tr>
<tr>
<td>Solution</td>
<td>0.44</td>
</tr>
<tr>
<td>Subgoals</td>
<td>0.42</td>
</tr>
<tr>
<td>Problem</td>
<td>0.42</td>
</tr>
<tr>
<td>Ends</td>
<td>0.41</td>
</tr>
<tr>
<td>Towers</td>
<td>0.38</td>
</tr>
<tr>
<td>Hanoi</td>
<td>0.38</td>
</tr>
<tr>
<td>Solving</td>
<td>0.38</td>
</tr>
<tr>
<td>Step</td>
<td>0.37</td>
</tr>
<tr>
<td>User</td>
<td>0.37</td>
</tr>
</tbody>
</table>

* No terms were removed and pseudodoc (log entropy weighting) was used.

The following “missing words” list was returned for the Problem Solving text. Missing words of some concern are decomposable/decompose/decomposition, and ignorable, irrecoverable, and reversible. These were terms that were important in the “Planning and problem decomposition” section of the text. As such, any inclusion of these words in the collaborative text signals or the participants’ summaries will not be used by LSA to compute similarity.

List of missing words:
achieves, advantageous, algorithms, anticipating, anticipating, astronomical, backtrack, brutal, commences, compatibility, consuming, corrections, deadend, decomposable, decompose, decomposition, dislike, effected, efficacy, emanating, exponentially, feasibility, fewest, fewest, finalizing, fine, fosters, generates, herbert, ignorable, instantiating, irrecoverable, metaphorical, minimized, modeling, modestly, notorious, paint, permeates, predicated, reapplied, recoverable, recursion, recursively, regrettable, restriction, restricts, reversible, roadblocks, roller, salesman, satisficing, shortest, slate, suffice, superlative, treasurer, undoing, undone, unduly, unlock, unpainted, virtues, wasted, workable.

The LSA spaces used in the analyses that follow (Research Question 2 and 3) seemed to adequately cover the content of the texts. The only concern was for the second text, for which there were a few important words that did not appear in the corpus used to create the semantic space. These words were found primarily in one section of text (out of four) and should still permit LSA to compute similarity fairly well.
A self-organising systems approach to social software

A note about the Latent Semantic Analyses

Hyphenated words often caused problems with the LSA analyses and so hyphens were removed from the documents submitted (texts, participant summaries, etc.). Quotations marks were also removed (double, single, and the possessive marker ‘s) for the same reason. The number of words reported in Tables 27 and 28 (see below) are for the texts as used in the LSA analyses, and as such, are necessarily slightly larger than the number of words reported above in the Methods section (since the removal of hyphens increases word count).

For the analyses below, similarity was computed using the ‘One-to-Many’ application. Document to document comparison (i.e., log entropy weighting) was used in all analyses.

Research Question 2 – Validity of the high signals

This section addresses the second research question;

2. How well do the stable collaborative signals reflect the main ideas in the text?
   2.1. How semantically similar, based on LSA will an artificial summary (consisting only of the high signals text), be to the original text?
   2.2. How does an artificial summary (consisting only of the high signals text) compare to a model summary?

LSA: high signals to the original text (RQ2.1)

For each text the high signals (after participant 40) were extracted and an “artificial summary” based on these words was created. These summaries were a collection of phrases in most cases and do not constitute a summary in the normal sense. Recall however, that LSA treats a document as a collection of words and does not recognise word order, sentence boundaries, or style. As such, it treats any summary as a sum of the words it contains. For this reason, it is possible to examine whether or not the collaborative text signals that emerged from the participants actions were, when taken as a whole, semantically similar to the overall text.

Flynn Effect text

There were 405 high signal words, 13.9% of the overall text, for the Flynn Effect text at the end of the study (see Table 27; also see Appendix B for the actual signal words
in context). The semantic similarity between this summary and the text was +0.78 (cosine).

Problem Solving text

There were 284 high signal words, 10.6% of the overall text, for the Problem Solving text at the end of the study (see Table 28, also see Appendix C for the actual signal words in context). The semantic similarity between this summary and the text was +0.80 (cosine).

Results Summary: Research Question 2.1

In both cases, the high signals summary had a very high semantic similarity to the text, even though fewer than 15% of the words in the texts were used in the summary.

Table 27: Semantic similarity between the Flynn Effect text and the author’s overview, and two artificial summaries

<table>
<thead>
<tr>
<th></th>
<th>Number of Words (Percentage of text)</th>
<th>Similarity: LSA cosine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flynn Effect text</td>
<td>2916</td>
<td>N/A</td>
</tr>
<tr>
<td>High Signals Summary</td>
<td>405 (13.9%)</td>
<td>+0.78</td>
</tr>
<tr>
<td>Author’s Overview as benchmark (first two paragraphs of the original text)</td>
<td>225 (7.7%)</td>
<td>+0.70</td>
</tr>
<tr>
<td>High Signals Summary: matched for overview length (&gt;=.80 importance score)</td>
<td>237 (8.1%)</td>
<td>+0.68</td>
</tr>
</tbody>
</table>
Table 28: Semantic similarity between the Problem Solving text and the author’s signals, and two artificial summaries

<table>
<thead>
<tr>
<th></th>
<th>Number of Words (Percentage of text)</th>
<th>Similarity: LSA cosine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem Solving text</td>
<td>2687 N/A</td>
<td></td>
</tr>
<tr>
<td>High Signals Summary</td>
<td>284 (10.6%)</td>
<td>+ 0.80</td>
</tr>
<tr>
<td>Author’s typographical text signals as benchmark</td>
<td>47 (1.7%)</td>
<td>+ 0.58</td>
</tr>
<tr>
<td>High Signals Summary: matched for number of author’s signals (&gt;=.90 importance score)</td>
<td>40 (1.5%)</td>
<td>+ 0.58</td>
</tr>
</tbody>
</table>

LSA: high signals compared to a model summary (RQ2.2)

For each text, the high signals artificial summary described above was compared to a model summary of the text. These models differed for the two texts and are described below.

Flynn Effect text

Author’s overview as a model summary.

For this text the author (Neisser, 1997) provided an overview, effectively a summary of the text (see Appendix H). Recall that this overview was removed from the text (the first two of three paragraphs) as used in the study, as such, participants did not read the overview. The author’s overview is treated as a useful benchmark. A shorter artificial summary was created matched to the length of the author’s overview (LSA is sensitive to text length). This was accomplished by retaining words achieving an importance score of .80 or higher. The overview and the shorter artificial summary were then compared to the text using LSA.

The similarity between the text and the author’s overview was +.70 (see Table 27 above). The second artificial summary – that was approximately the same length as the author’s overview – was almost as similar to the text with a cosine of +.68. These results indicate that the collaborative text signals that emerged from the participants’ actions
captured the semantic content of the overall text almost as well as the author’s overview of the text.

**Problem Solving text**

*Author’s typographical text signals as a model summary.*

Unfortunately no adequate overview was provided by the author that could be used as a benchmark as with the previous text (there was only a general introduction to the chapter). However, in this text the author (Holyoak, 1990) provided many typographical text signals through the use of italics. Recall that these were removed from the text as used in the study, as such, they did not influence the participants. There were 47 such words (with hyphens removed) that were treated as a very short “summary” of the test. A second, shorter artificial summary was created matched to the number of words that the author signalled (LSA is sensitive to text length). This was accomplished by retaining words achieving an importance score of .90 or higher. The author’s signals are treated as a useful benchmark.

The similarity between the text and the author’s signals was +.58 (see Table 28 above). The second artificial summary – that was approximately the same length as the author’s signals – had the same similarity to the text. These results indicate that the collaborative text signals that emerged from the participants’ actions captured the semantic content of the overall text as well as the author’s signals of the text.

*Author’s typographical signals compared to the collaborative typographical signals.*

A second analysis was performed to investigate whether or not the text signals that emerged from the indirect collaboration of the students matched those by the author (Holyoak, 1990). This analysis did not rely on LSA. In all, there were 30 one or two word phrases signalled by the author, consisting of 45 words. Hyphenated words counted as a single word, which is how CoREAD treats hyphenated words for highlighting and signalling purposes (for the author’s signals see Appendix C, note 4). Of these 45 words, how many were collaboratively signalled – at the high (red) level – by the group? There were 31 matches out of the 45 cases (69%) whereby the words signalled by the author
were also independently signalled by the group. Please note that matches were for the exact word in the text by position, not simply the same word occurring anywhere in the text. The likelihood of this happening due to random chance is incredibly small. Only 10.5% of the text reached the high signal (red) state at the end of the study. The probability that 31 of the 45 words signalled by the author would also be in a high signal state can be calculated using the binomial theorem; it is $1.60 \times 10^{-20} \%$. The most likely chance outcomes are 2 to 7 matches (i.e., successes) which account for over 85% of the possibilities (see Figure 14). There is less than a 1% chance of more than 10 matches. This shows that the typographical text signals that emerged matched the author’s signals far more often than could be explained by chance. The indirect collaboration of the participants seems quite capable of yielding text signals that capture what an author might include. It must be noted of course that the high signals numbered 284 words and so include far more information than the author’s signals (45).

![Figure 14: Binomial probability distribution for 45 trials when the probability of a match at each trial is 0.105](image-url)
Results Summary: Research Question 2.2

For both texts the collaborative signals performed as well as the model summary. The high text signals that emerged from the participants’ indirect collaboration seem to provide reasonably good semantic coverage of the text.

Results Summary: Research Question 2

The high text signals that emerged from the collaborative efforts of the group appear to capture the semantic content of the texts quite well. This was true even when the high signals were restricted in length to match a model summary. In this case, the high signals summaries performed as well as the model summaries. That is, they were as semantically similar to the text as the model summaries were.

Research Question 3 – Better summaries post stability?

This section addresses the third research question:

3. Do readers produce better summaries when the text has stable collaborative signals?
   Dividing the readers into pre-stability and post-stability groups:
   3.1. Are the post-stability summaries more semantically similar, based on LSA, to the original text?
   3.2. How do the post-stability summaries compare to a model summary?

In the results for the first research question it was determined that the signals stabilised following the 30th participant. For the analyses that follow the sample was separated into quartiles by position in the sequence (participants 1-10, 11-20, 21-30, and 31-40) in order to compare the three pre-stability groups to the post-stability group (i.e., the fourth quartile).

Are post-stability summaries more semantically similar to the original text? (RQ3.1)

The participants’ summaries were compared to the relevant text using LSA. The summary statistics for the texts are presented in Table 29 below. Clearly, these data do not suggest that the post-stability group produced summaries that were more semantically similar to the text. The overall pattern may also be seen across all participants in Figure 15 below.
Table 29: Similarity of the participants’ summaries to the texts by quartiles (n=10); measured by LSA cosine, $M$ and $SD$ reported.

<table>
<thead>
<tr>
<th>Quartiles</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Flynn\text{ Effect text}$</td>
<td>$M$</td>
<td>+0.75</td>
<td>+0.76</td>
<td>+0.74</td>
</tr>
<tr>
<td></td>
<td>$SD$</td>
<td>0.06</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>$Problem\text{ Solving text}$</td>
<td>$M$</td>
<td>+0.81</td>
<td>+0.81</td>
<td>+0.80</td>
</tr>
<tr>
<td></td>
<td>$SD$</td>
<td>0.03</td>
<td>0.04</td>
<td>0.02</td>
</tr>
</tbody>
</table>

How do the post-stability summaries compare to a model summary? (RQ3.2)

For each text, the mean cosine for the post-stability group was compared to the cosine for the model summary described earlier (see Results – Research Question 2.2). These are measures of the semantic similarity between the participants’ summaries and the text and the model summary and the text, respectively. Figure 15 (below) shows the cosines for each participant and the model summary (the horizontal line). As the graph makes clear, some participants did produce summaries that were more semantically similar to the text (higher cosine) than the model summary. However, the post-stability group (participants 31-40) did not outperform the model summary on average. The relevant statistics are presented in Table 30 below.

Table 30: Comparison between the post-stability participants’ summaries and the model summary, for each text (LSA cosine measure)

<table>
<thead>
<tr>
<th></th>
<th>Post-stability summaries, $M$ ($SD$)</th>
<th>Model Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Flynn\text{ Effect text}$</td>
<td>+0.74 (0.05)</td>
<td>+0.78</td>
</tr>
<tr>
<td>$Problem\text{ Solving text}$</td>
<td>+0.80 (0.03)</td>
<td>+0.80</td>
</tr>
</tbody>
</table>

* Note: participants 31-40, n=10

Results Summary: Research Question 3

These results do not suggest that the stable collaborative text signals assist readers to produce better summaries. The study as designed was not powerful with respect to the third research question. Since the sample needed to be split based on the point when the signals stabilised (i.e., after participant 30) only 10 participants were included in the post-stability group. The size of this group is much too small given the variability among the participants with respect to the quality of their summaries. It may have been too
ambitious to attempt to address all three research questions with one study. Future research will be required to determine if the collaborative text signals assist readers as has occurred in traditional signalling research (Lorch, 1989; Lorch et al., 1995). It is possible that readers do not respond to collaborative text signals in the same way as authorial text signals.
Figure 15: Semantic similarity between 1) the participants’ summaries, and 2) the model summary (indicated with a horizontal line) and the relevant text
CHAPTER 6: DISCUSSION

In this chapter the results are discussed with respect to the three hypotheses. As well, I examine whether CoREAD was adaptable, flexible, and robust. Two potential problems affecting the use of CoREAD are also examined. One problem addresses a concern educators might have regarding the use of CoREAD and similar systems. Last, ideas for future work with CoREAD are outlined.

Stable, valid, and useful for summarisation?

Hypothesis 1: Stability

The use of CoREAD by a group of heterogeneous readers will eventually lead to stable typographical text signals developed in a self-organised fashion.

Recall that an emergent phenomenon is one that is a stable pattern, occurs at a level above the agents themselves, and should not be directly predictable from agent level behaviour. The difficulty in assessing whether emergence has occurred arises from the difficulty in assessing stability, particularly since complex systems are generally undergoing constant change and may stabilise more than once at different times and in different configurations.

Stability was assessed using three methods and all showed that the high signals had reached a fairly high level of stability after the 30th participant, for both texts. This result is somewhat surprising in that the texts were quite different in content and genre. Both texts were expository but the Flynn Effect text presented arguments and evidence in the last two sections, whereas the Problem Solving text was purely informational.

Nor was there an information cascade, participants added “new” opinions to the history throughout the study.

55 Both texts were expository but the Flynn Effect text presented arguments and evidence in the last two sections, whereas the Problem Solving text was purely informational.
Hypothesis 2: Validity

The collaborative signals indicating highly important text sections (i.e., red font) will tend to correspond with the semantic meaning of the text.

The high signals that emerged did correspond with the main ideas in the text for the most part. This was indicated by the fact that the high signal words taken together were as semantically similar to the text as model summaries, once length of summary was controlled for. Also, for the Problem Solving text, the correspondence between the author’s signals and the collaborative text signals was well in excess of anything that could be explained by chance. Since authorial signals are considered useful additions to a text then these collaborative text signals must also be considered useful until evidence suggests otherwise.

Hypothesis 3: Effects on summarisation

The presence of stable collaborative signals in a text will positively influence summarisation. Students from the post-stability period will produce better summaries than those students who read the text before the signals have stabilised.

As mentioned in the results, a properly controlled study will need to be conducted to determine if the participants were influenced by the text signals. The results from this study show that the direction and level of agreement between the actions taken by participants (i.e., highlighting text sections) and the text signals they viewed was consistent with an influencing effect. Participants agreed with the high signals by also highlighting those text sections 77% and 73% of the time, for the two texts respectively. This is not consistent with an information cascade (Surowiecki, 2004) or narrowing (Clark, 2003, p.182-3), however, since disagreements were still common across the participants.

The study did not provide evidence of improved summary writing performance for the post-stability group. The sample size (n=10) for this group was quite small. More importantly, the summaries were probably affected by a large number of variables, such as differences in summary writing ability. Although the summarisation task provided a realistic context for using the highlighting function it may not be as useful as a measure of the impact of the text signals. Indeed, most signalling studies have used recall measures
A self-organising systems approach to social software

(Lorch, 1989). The reliability of many recall measures (e.g., short answer and multiple choice questions) is in general much higher as well, thereby lowering the random error associated with the measure. This would increase the chances of obtaining statistically significant results.

As previously explained, the summary task was used in large part to provide the participants with a realistic context in which to select important passages and highlight them. However, the summary task is probably a poor choice for evaluating differences without employing controls for summary writing ability (A. L. Brown & Day, 1983; Winograd, 1984). Some form of matching by summary writing ability (e.g., a within-group design or matched samples) should be employed; perhaps using two texts where one is read with stable signals and the other with no signals.

In addition, the signals used in traditional text signalling research are likely to have somewhat different effects on participants because of the origin of the signals. First, the signals may be more trusted because participants believe an authority figure, the author or researcher, has placed the signals in the text. Second, particularly in immediate cued recall tasks, the participants may assume that the questions they will be asked will undoubtedly focus on the signalled information. In the case of CoREAD, the signals are not from an authoritative source, but other students. As well, the participants have no reason to think that their summaries will be assessed based on the signals that they saw.

Given the low power – because of the small sample size for this research question – it is not possible to conclude that collaborative signals do not have an effect. One question that should be asked, and researched, is whether identical signals when presented to participants as “from the author” yield different results from when they are presented as “collaboratively generated”. If the authorial signals were found to have positive effects (as in previous research) but not the collaborative signals then it would suggest that readers do not critically evaluate the signals from trusted sources. Rather it would suggest that readers accept the importance of these authorial signals with little thought. This might occur since preparing for testing is paramount and the content signalled by the author can be assumed to have a higher probability of appearing on the test. If this is the case, it may be a cause of some concern. First, there is no research on the quality of the text signals that are included in actual educational texts. Second, there are no guidelines
or objective techniques for selecting which text segments to signal for emphasis, and this would depend to some degree on the reader’s assigned task (also see Lemarie et al., 2008). Third, outside of formal educational settings the inclusion of text signalling by authors may be more limited. If students come to rely on such signals to recognise important text segments for additional or special processing then what do students do when reading non-educational texts?

**Summary of stable, valid, and useful text signals?**

The results indicated that the participants were able to generate stable text signals after approximately 30 of them had read a text. Additionally, the high text signals corresponded well with the semantic content of the text. The last hypothesis, however, was not supported by the evidence. The participants may not have responded to the collaborative text signals as they would have if an author or instructor had prepared the signals. Sample size issues when addressing this hypothesis also resulted in low power – though the basic effect size (i.e., the mean differences) was very small in any case. Future research will need to examine why these signals did not seem to improve performance on the summarisation task.

**Adaptable, flexible, and robust?**

Recall that a self-organising systems approach was selected because they tend to be adaptable, flexible, and robust. They are adaptable since solutions to problems emerge over time in response to changes in the context of the problem or the available solutions. They are flexible because these systems can solve a class of problems and can respond to a variety of specific situations. Finally, they are robust because removing or changing some of the individual agents in the system (e.g., adding a reader with poor skills or knowledge) generally does not compromise the performance of the system overall (the collective). In the following I address whether CoREAD is likely to produce adaptable, flexible, and robust text signals when used by a group of students.

**Adaptable?**

One could argue that the collaborative text signals became adapted to the summarisation task demands because they stabilised and were semantically valid. Even
though at this time it does not appear that the signals, however stable and valid, assisted
the participants to write better summaries. But how adaptive is the system? Could text
signals move out of a current stable state into a new stable state if there was a change in
the readers, task, or broader educational context? For example, what if an instructor
discussed a common student error in class and directed students to re-read a particular
text (or part thereof). What if students were given a new and different task to complete
with a previously read text (e.g., critique the article or compare it with another article)?
Or perhaps texts are carried over from year-to-year with the collaborative signals of
previous years intact (this may or may not be a wise choice). How would the signals
adapt to these new circumstances? It is entirely likely that the text signals would adapt in
these situations given the new information, new task demands, or new readers. But this
adaptation may occur more slowly if a stable point had already been reached. This will
have to be examined in future work.

The weighting function may provide a means to tune the speed at which the text
adapts. The weighting function could be adjusted to promote adaptation from a currently
stable state. For instance, the weighting function could be changed so that earlier
contributions in the history had even less impact on the importance scores for a period of
time (until returning to the usual function). This would further minimise the direct effects
of the earliest contributions in the history on the importance scores. Effectively, this
would promote more rapid changes in the signals by favouring new opinions, allowing
the readers to find a new point of stability better adapted to the new context.

A much simpler approach might be to give readers the ability to turn signalling on
and off. Simply ignoring the signals that had emerged under different task conditions
may have the same effect. This might be a positive, unintended consequence of giving
readers more control over the software.

Flexible?

Although the two texts were both on topics from psychology they were about
different main concepts. As well, the participants reported knowing far less about the
Problem Solving text before having read it. Yet, in both cases the participants generated

56 CoREAD already supports this function but it was not available to the participants in the study.
stable and valid signals for the texts. This suggests that CoREAD, when used by a group of diverse readers who collectively have some relevant prior knowledge\(^{57}\), is flexible enough to generate appropriate text signals for a range of texts.

*Robust?*

As has been shown, a heterogeneous group of readers with very few members who were well informed about the main topics of the texts were still able to produce relevant signals. The self-organising system of participants using CoREAD can obviously tolerate some poor contributions to the collective history. As such, it seems to be fairly robust, as would be expected of a self-organising system.

*Summary of adaptable, flexible, and robust?*

CoREAD seems to produce text signals in a manner that is flexible and robust. The degree to which the signals could adapt to changes in the task or the participants involved (e.g., a new class of students who have “inherited” a text) remains to be studied. This could be examined by changing a reading task (e.g., prepare a critique instead of a summary) once a subset of the participants had already read a text. This would reveal how adaptable the signals are to such changes. Modifications to CoREAD’s weighting function might be necessary to encourage more rapid adaptation in such cases.

*CoREAD: Two potential problems*

Two potential problems affecting the use of CoREAD, and social software more generally, in classrooms are known as the “free rider” and “cold start” problems. These are examined below.

*The “free rider” problem*

The “free rider” problem\(^{58}\) is one where some students simply use others to accomplish their assigned learning tasks. This would be seen in group work, for example,

---

\(^{57}\) Including relevant conceptual knowledge, knowledge of appropriate reading strategies, and knowledge of the text genre.

\(^{58}\) A similar concern regarding social software for educational purposes are “lurkers” (Ebner & Holzinger, 2005), non-contributing members who still make use of and benefit from the available information. Whether lurkers are a problem or simply a form of “legitimate peripheral participant” is debatable.
where some students rely on their group members and contribute little themselves. A similar problem could occur with tools like CoREAD. Some students could wait until a text had been read by a number of others and simply accept the resulting text signals without carefully reading the text. As such, they would also not make any “independent” contributions to the collaborative history. If this became the norm then an information cascade would occur.

Although possible, the reading task is a means to an end, namely understanding the text (or writing a summary, passing a test, etc.). Each individual reader is separately evaluated later and so he should be motivated to evaluate the text plus collaborative signals, himself. Any benefits to the reader will be decided by a process outside of the self-organising system generating the signals. This is unlike stock markets, which can often be victims of information cascades because during a boom merely following along (i.e., buying stock) is itself rewarding. Surowiecki’s (2004) concern about groups of people making decisions in a sequence with a lack of independence between the members, follows from this observation. However, an example of farmers adopting a new crop (Surowiecki, 2004, p.61-2) shows that they were far more willing to combine their own experimentation with the group’s suggestion (i.e., adopt the new crop). An information cascade did not occur. I would argue that this is another example of how separating the reward context from the system generating ideas or solutions minimises the potential of an information cascade with all of its potential negative consequences.

Results from this study have indicated that the participants continued to add their own contributions – by disagreeing with the current collaborative signals – regardless of the participant’s position in the sequence. Stable signals emerged, but not because participants later in the sequence stopped contributing. Although this finding does not rule out the free rider problem it is certainly reassuring.

In conclusion, the “free rider” problem may be less likely to occur with CoREAD than one might imagine. However, educators also need to question under what circumstances obtaining help from a group is legitimate. Students with less relevant prior knowledge or weaker reading comprehension skills might well benefit from reading a text after it had been augmented with collaborative text signals. If reading a text with authorial
signals is legitimate then why should reading a text with collaborative text signals not be legitimate as well?

*The “cold start” problem*

There is a possibility of a “cold start” problem that often affects social software (Dron, 2005; Dron et al., 2001; Vassileva & Sun, 2007). Social software needs to be used by a group of people before anything useful from the group’s activity can emerge. At first, there is no benefit to using the software since there is no “community wisdom” available yet. This lack of incentive to use the software initially can make it difficult for the system to acquire enough users, or enough content in some cases as well, for it to become useful.

In many social software systems users must enter a review, register a vote (e.g., clicking “I like/don’t like it” on StumbleUpon), or tag or categorise a web site, for example. These activities are sometimes in addition to any main task the user is engaged in. In such cases, the software designers may include reward structures in the software to promote activity (Vassileva et al., 1999). Otherwise, one must hope that some initial group of users contributes out of a sense of altruism.

CoREAD should not suffer from the cold start problem as much as other social software. There are two reasons for this. First, in formal educational contexts students are required to read certain texts and would typically do so for their own benefit with or without CoREAD. Second, there is no additional work required from the reader. The readers highlight the text for their own use. This highlighting information is then collected passively (Wexelblat & Maes, 1999), it is an implicit source of information from the readers (Recker et al., 2003). Therefore, the collaborative text signals are merely a by-product of individual work.

*Summary of two potential problems*

CoREAD is less likely to suffer from the free rider problem because the reading context that generates the signals is separate from the assessment context. Each student will be assessed individually and should be motivated to evaluate the collaborative text signals before accepting them. As well, students with different prior knowledge would
A self-organising systems approach to social software

benefit from attending to different parts of a text and so no set of text signals is “correct” for every student.

Similarly, the cold start problem should not be a serious concern since students are assigned texts to read and will be at a disadvantage if they do not read the texts. As well, they will incur no additional cost from using CoREAD since the text signals result as a by-product of their own work.

**Future Work**

This study provides the groundwork for future research into collaborative text signalling. CoREAD seems capable of generating stable and potentially useful text signals as a by-product of individual work. Evidence that collaborative text signals are beneficial to readers has not yet been found. This may have resulted because readers respond differently to collaborative signals than they do to authorial signals. Additionally, the task used in the study was a summarisation task with measures of performance based on semantic similarity (i.e., using LSA). Most of the previous studies on text signalling have asked participants to read a text knowing they will be tested immediately thereafter. Measures of performance were tests of recall (free or cued). As a first step, it will be necessary to address two questions.

1) How do readers respond to authorial versus collaborative text signals? and,
2) What are the effects of collaborative text signals on recall, comprehension, and the execution of specific tasks (e.g., writing a critique, writing a summary)?

Future work should also examine how CoREAD is used by learners in real educational settings. How would learners sequence themselves? Would “stronger” students elect to read early in the history knowing that signalling would not be available (thereby overcoming a “cold start” problem)? Would “weaker” students wait to gain from the knowledge of the others? If given the ability to turn the signals on and off with a toggle switch (CoREAD already supports this function) how would readers use it? Would more individualistic learners keep the signals off? Would they turn the signals on if they became confused or uncertain?

Finally, it would be useful to have an agent-based model (Goldstone et al., 2008; Miller & Page, 2007; Page, 2007) of learners using CoREAD. This would allow
researchers to “experiment” with parameters that might affect performance. These include:

1) the CoREAD weighting function
2) the number, characteristics (prior knowledge, skills, attitudes), and order of the readers in the group
3) the minimum importance score required (i.e., cut-offs) for each text signal state
4) user control over the signalling (ability to turn it on and off)

A model would be very useful since collecting empirical data on the immense variety of conditions, and combinations of conditions, is not feasible. Research evidence often does not permit predictive calculations of parameter values. For example, Lorch (1995) found signals useful if they were not too numerous. However, the study used a dichotomous independent variable with 5% versus 50% of the words signalled. This shows that excessive signalling results in no benefits but does not answer the question, “how much signalling is too much?” Also, how does the amount of signalling interact with other conditions? Even very simple agent-based models that accurately capture only some aspects of the behaviour would be greatly beneficial. Results from simulations could potentially narrow the field of ideal parameters for empirical testing with actual participants.
CHAPTER 7: CONCLUSION

Humans create their cognitive powers by creating the environments in which they exercise those powers.

Hutchins (1995a, p. xvi)

The CoREAD project has demonstrated that readers can produce stable and valid text signals through a self-organising process. This is both a surprising and modest success. Surprising since these results occurred without any of the normal human interaction that one thinks of when people collaborate. Modest since there was no evidence that the collaborative text signals assisted the readers to produce better summaries.

As Pea (1993) pointed out the distributed perspective raises the issue of trade-offs. For example, allowing students to use artifacts may permit more individuals to participate in activities because some aspect of the cognitive activity is changed or removed through the use of the artifact. For example, using a computer may permit students to rapidly construct graphs without having to learn techniques for hand crafting graphs. The trade-off is that the students do not learn how to create graphs by hand. In some cases, such as this one, what is lost is insignificant given that modern professional approaches also use the same computer tools available. In other cases, the trade-off may be more contentious.

The use of social software like CoREAD is likely to raise discussions of such trade-offs. Will some students, by using the software, become less able to assess a text and locate the main ideas themselves? Like many such concerns, empirical data are the best way in which to answer the basic question. Once the data are in, however, the relative value one places on the advantage attained and the skill or knowledge lost will still affect the decision made. Are the benefits worth the costs? Kirsh (2006) raises this idea with respect to metrics of performance. For example, if an artifact is introduced into a cognitive activity how does performance change with respect to time spent, errors made, the ease of learning, etc. Many metrics could be used to examine how CoREAD affects performance. These could include:

- reading time
- recall measures
- quality of the work based on the readings (essays etc.)
how soon students complete the readings
• performance with respect to prior knowledge
• changes in personal annotation activities

Ultimately, finding that students come to depend in some sense on CoREAD would not necessarily be a negative outcome. The trade-off might not be such that the costs outweighed the benefits.

People today are already dependent on software for many tasks. Software is used to “remember” e-mail addresses of one’s many colleagues, to organise and locate useful references (e.g., Endnote), to navigate the streets and highways of unfamiliar locations using GPS devices, and locate information on the web. Creating artifacts and then depending on them is quite normal, and as Clark suggests, perhaps a feature of the human mind (Clark, 2003).

CoREAD merely makes one aware of the thoughts of others. Being aware of and referring to the ideas of others is perhaps the most common feature of scholarly work. Although objections to the use of social software are potentially valid, we should perhaps base our discussions on some sensible empirical accounting of the real costs and benefits. Unlike far too many of the disputes in education, one should avoid contrasting imagined costs against improperly measured benefits.

Social software is part of the online experience for many people, and is potentially a powerful means of large-scale collaboration in educational contexts. Although there may indeed be costs involved, many current uses of social software and the results of this study suggest that the benefits of mass collaboration may outweigh the costs.

Statement of originality

The design of CoREAD was greatly inspired by the previous work on history-enriched digital objects (Hill & Hollan, 1993; Hill et al., 1992; Hollan et al., 2000). Unlike the previous work that was based on a metaphor of wear, CoREAD is based explicitly on complex, self-organising systems. As such, CoREAD not only maintains a history-of-use but aggregates the history in ways that attempt to foster positive and negative feedback loops, balancing exploration and exploitation.
The work presented makes three contributions to the literature. First, I have presented a theoretical synthesis of complex systems, social software, text signalling and distributed cognition. As previously noted, others have provided syntheses of complex systems and social software (Dron, 2007b; Dron et al., 2000) and complex systems and distributed cognition (Gureckis & Goldstone, 2006; Heylighen et al., 2004; Poirier & Chicoisne, 2006). The synthesis presented leads to the conclusion that social software may produce history-enriched digital artifacts, which are a “new” form of distributed cognition: New because they involve large-scale indirect, stigmergic interaction. These stigmergic artifacts should be investigated and designed informed by an understanding of complex systems. Second, the design of CoREAD provides a platform for investigating how readers collaborate indirectly to generate self-organised text signals. Although CoREAD now only supports typographical text signals other types of signals such as overviews, previews, summaries, and marginalia (anchored notes) could also be added. Third, the study has demonstrated that readers can indeed generate stable signals in a self-organising fashion. Additionally, these signals seem as valid as authorial signals.
REFERENCES


A self-organising systems approach to social software


A self-organising systems approach to social software


A self-organising systems approach to social software


A self-organising systems approach to social software


APPENDIX A: RESEARCH ETHICS BOARD APPROVAL

McGill University
Research Ethics Board Office
1555 Peel Street, 11th floor
Montreal, QC H3A 3L8
Tel: (514) 398-6831
Fax: (514) 398-4644
Ethics website: www.mcgill.ca/researchoffice/compliance/human/

Research Ethics Board II
Certificate of Ethical Acceptability of Research Involving Humans

REB File #: 294-0308

Project Title: Self-organised, collaborative text signalling

Principal Investigator: Andrew Chiarella
Department: Educational & Counselling Psychology

Status: Ph.D. student
Supervisor: Prof. Susanne Lajoie

Funding Agency and Title (if applicable): N/A

This project was reviewed on April 7, 2008 by Expedited Review


Mark Baldwin, Ph.D.
Chair, REB II

Approval Period: April 7, 2008 to April 6, 2009

This project was reviewed and approved in accordance with the requirements of the McGill University Policy on the Ethical Conduct of Research Involving Human Subjects and with the Tri-Council Policy Statement: Ethical Conduct For Research Involving Humans.

* All research involving human subjects requires review on an annual basis. A Request for Renewal form should be submitted 3-4 weeks before the above expiry date.
* When a project has been completed or terminated a Final Report form must be submitted.
* Should any modification or other unanticipated development occur before the next required review, the REB must be informed and any modification can't be initiated until approval is received.
APPENDIX B: FLYNN EFFECT TEXT WITH SIGNALS

Notes:
1. The signals are indicated using the same colour coding as used in CoREAD and are from the end of the study, after participant 40.
2. There are 23 CoREAD screens (title plus 22 paragraphs) with each screen enclosed in a box.
3. The paragraphs are reproduced as they were viewed by the participants.

Because there are many different forms of mental ability, there are also many different kinds of tests and test items. Some are verbal and others are visual in format. Some tests consist only of abstract-reasoning problems, and others focus on such special competencies as arithmetic, spatial imagery, reading, vocabulary, memory or general knowledge. The broad-spectrum tests, which establish actual IQ scores, typically include a wide variety of items. Before considering these general instruments, however, we must take a brief look at the relations among different specialized tests and how those relations are traditionally interpreted.

The degree to which any two tests measure something in common can be indexed by their correlation r, which in principle ranges from -1 to +1. A positive r means that individuals who score high on one test also tend to score high on the other; a negative r, which rarely occurs in this context, means that high scores on one test go with low scores on the other. When the same group of individuals takes a number of different tests, one can compute an r for each pair of tests considered separately, and the result is a correlation matrix. For intelligence tests, the correlation matrix tends to consist of r's that are all positive, but well below 1.

Early in this century, the British psychologist Charles Spearman made the first formal factor analyses of such correlation matrices. He concluded that a single common factor accounted for the positive correlations among tests - a notion still accepted in principle by many psychometricians. Spearman christened it g for "general factor". In any test battery, the test that best measures g is - by definition - the one that has the highest correlations with all the others. The fact that most of these g-loaded tests typically involve some form of abstract reasoning led Spearman and his successors to regard g as the real and perhaps genetically determined essence of intelligence.
Although that view remains widely held, it is not a necessary conclusion. Other factor analyses of such data are possible and have been proposed. Today, some psychometricians regard \( g \) as little more than a statistical artifact, whereas others seem even more convinced than Spearman himself that it reflects a basic property of the brain. Whatever \( g \) may be, at least we know how to measure it. The accepted best measure is a (usually untimed) test of visual reasoning called Raven's Progressive Matrices, which was first published in 1938 by Spearman's student John Raven and is now available in several different levels of difficulty. As we shall see, Raven's test plays a central role in recent analyses of the worldwide rise in test scores.

In contrast to specialized instruments like the Raven, the tests most widely used in America include a wide variety of different items and subtests. The best known of these "IQ tests" are the Stanford-Binet and the various Wechsler scales. The Wechsler Intelligence Scale for Children (WISC), for example, has five "verbal" subtests (information, comprehension, arithmetic, vocabulary and explaining similarities) and five "performance" subtests in which a child must copy designs using patterned blocks, put several related pictures in their proper order and so on. A child's scores on these subtests are added up, and the tester converts the total to an IQ by noting where it falls in the established distribution of WISC scores for the appropriate age.

That distribution itself - the crucial reference for assigning IQ scores - is simply the empirical result that was obtained when the test was initially standardized. By convention, the mean of each age group in the standardization sample defines an IQ score of 100; by further convention, the standard deviation of the sample defines 15 IQ points. Given appropriate sampling and a normal distribution, this implies that about two-thirds of the population in any given age group will have IQs between 85 and 115.

IQ defined in this way reflects relative standing in an age group, not absolute achievement. The mean-scoring eight-year-old attains a higher raw score on the WISC than the mean-scoring seven-year-old, but both have IQs of 100 because they are at the middle of their distributions. So in one sense (as measured by raw scores), a normal child becomes systematically more intelligent with age; in another sense, his or her intelligence remains relatively stable. Although raw scores rise systematically throughout the school years, IQs themselves rarely change much after age 5 or 6.
IQ tests do not remain fixed forever. Most of the major ones have been updated from time to time. The 1949 WISC, for example, was superseded by the WISC-R in 1974 and by the WISC-III in 1991. The revised versions are standardized on new samples and scored with respect to those samples alone, so the only way to compare the difficulty of two versions of a test is to conduct a separate study in which the same subjects take both versions. Many such studies have been carried out, and James Flynn, a political scientist at the University of Otago in New Zealand, summarized their results in 1984. In virtually every instance, the subjects achieved higher scores on the older version of a test. For example, in David Wechsler's own study of his revised adult test—the Wechsler Adult Intelligence Scale-Revised (WAIS-R)—a group of subjects who averaged 103.8 on the new WAIS-R had a mean of 111.3 on the older WAIS. This implies that the actual IQ-test performance of adults rose by 7.5 points between 1953 (when the old WAIS was standardized) and 1978 (when the WAIS-R was standardized), which is a rate of about 0.3 IQ points per year.

These gains are not limited to the WAIS, to adults or to the United States. In an influential series of papers, Flynn showed that the increasing raw scores appear on every major test, in every age range and in every modern industrialized country. The rise itself is now often called "the Flynn effect". The increase has been continuous and roughly linear from the earliest days of testing to the present. On broad-spectrum tests such as the WISC and the WAIS, Americans have gained about 3 IQ points per decade, or 15 points over a 50-year period. It is interesting to compare this total with the much-discussed gap between the mean test scores of Caucasian and African Americans, which is also about 15 points (one standard deviation of the IQ distribution). Given that the IQ of the population as a whole has increased by a similar amount just since the 1940s, that gap does not seem so large.

The pattern of score increases for different types of tests is somewhat surprising. Because children attend school longer now and have become much more familiar with the testing of school-related material, one might expect the greatest gains to occur on such content-related tests as vocabulary, arithmetic or general information. Just the opposite is the case—"Crystallized" abilities such as these have experienced relatively small gains and even occasional declines over the years. The largest Flynn effects appear instead on highly g-loaded tests such as Raven's Progressive Matrices. This test is very popular in Europe; the Dutch data mentioned earlier came from a 40-item version of Raven's test. Using the 1952 mean to define a base of 100, Flynn has calculated average Dutch Raven IQs for subsequent years. The mean in 1982 was 121.10—a gain of 21 points in only 30 years, or about seven points per decade. Data from a dozen other countries show similar trends, which seem to be continuing into the 1990s. Whatever g may be, scores on tests that measure it best are going up at twice the rate of broad-spectrum tests like the WISC and WAIS, while the tests most closely linked to school content show the smallest gains of all.
These gains are far too rapid to result from genetic changes. There evidently are substantial environmental influences on $g$, even if we do not clearly understand them at the present time. Moreover, the sheer size of the gains undermines the very concept of "general intelligence". To see why, consider that individuals with IQ scores over 130 are typically regarded as "very superior" and those with scores under 70 as "intellectually deficient". In any normal distribution of scores with a mean of 100 and a standard deviation of 15, about 2.25 percent of the population can be expected to fall into each of those categories. This was indeed roughly the case for the first Stanford-Binet standardization sample in 1932. An ongoing rise of 0.3 IQ points per year means, however, that if a representative sample of the American children of 1997 were to take that 1932 test, their average IQ would come out to be around 120. This would put about one-quarter of the 1997 population above the 130 cutoff for "very superior" - 10 times as many as in 1932. Does that seem plausible?

If we go by more recent norms instead, we arrive at an equally bizarre conclusion. Judging the American children of 1932 by today's standards - considering the scores they would have obtained if they had somehow taken a test normalized this year - we find that their average IQ would have been only about 80! Hardly any of them would have scored "very superior," but nearly one-quarter would have appeared to be "deficient". Taking the tests at face value, we face one of two possible conclusions - either America is now a nation of shining intellects, or it was then a nation of dolts.

If we focus instead on the most $g$-loaded tests, the gains seem even more preposterous. The mean Raven IQ in the Netherlands rose by 21 points between 1952 and 1982; extrapolating backward, there has probably been something like a 35-point increase since the 1930s. Dutch 19-year-olds today are getting scores that would have been more than two standard deviations above the mean in their grandfathers' time. The size of these gains boggles the mind. If they do not reflect some trivial artifact, we (and especially the Dutch) must be living in a truly remarkable age of genius. As Flynn puts it, the data imply that dozens of nations should now be in the midst of "a cultural renaissance too great to be overlooked". Because that does not seem to be happening, Flynn concludes that the tests do not measure intelligence but only a minor sort of "abstract problem-solving ability" with little practical significance.

Flynn's extreme position can help us organize the range of hypotheses that have been proposed to explain the rise in scores. Some hypotheses - including increases in test-taking sophistication and in the motivation to score well - are consistent with the view that there has been no real rise in intelligence at all. Other possibilities - including the impact of worldwide improvements in health and nutrition - conflict with Flynn's conclusion, suggesting that intelligence has really gone up. Several more subtle hypotheses - that the gains have been produced by changes in schooling, in child-rearing practices or by more general aspects of culture - lie in between these two extremes. Let us consider two of these possibilities.
In the countries where IQ scores are rising, people are also staying in school much longer than their parents and grandparents did. Could sheer duration of schooling be responsible for the gains? This hypothesis is plausible because the general effect of schooling on IQ is very well established. Many studies indicate that children who do not attend school for one reason or another score lower on the tests than their regularly attending peers. One especially unfortunate example of that principle appeared in the 1960s, when some Virginia counties closed their public schools to avoid racial integration. Compensatory private schooling was available only for white children. On average, the African-American children who received no formal education during that period fell back at a rate of about six IQ points per year.

Fortunately, such episodes are rare. School attendance through the elementary grades is virtually universal in all modern industrialized countries. In the United States, for example, more than three-quarters of the population goes on to finish high school. This universality makes it difficult to separate the contributions of age and schooling to mental development, because the “average eight-year-old” is an eight-year-old who has been in school for two or three years. Some separation is possible, however, because admission to first grade in most school systems is governed by an arbitrary age cutoff, such as six years old by September 1 of a given year. This practice ensures that the children in any given grade will vary in age by up to a year and that there will be children in different grades who are very nearly the same age.

In 1987, Sorel Cahan and Nora Cohen of the Hebrew University in Jerusalem took advantage of these birth-date distributions in an ingeniously designed study. They administered 12 different brief tests - presented together in a single session - to some 10,000 fourth-, fifth- and sixth-grade children in the Jerusalem schools. Using a complex statistical analysis based on birth dates and school admission, Cahan and Cohen compared the effects of a year of school (controlling for age) with those of a year of age (controlling for school) on each test separately. As one might expect, schooling mattered more than age for every test of verbal or numerical skills. More surprising, perhaps, is that schooling also made a contribution - albeit smaller - to performance on several nonverbal tests of abstract and visual reasoning. One of those nonverbal tests consisted entirely of items from the Raven.

Despite those data, schooling is not an altogether satisfactory explanation of the secular rise in test scores. For one thing, elementary school children tested with the WISC show gains comparable to those of adults who take the WAIS. Since elementary education was already universal in the 1930s, the WISC gains cannot be attributed to increased years of schooling. Another argument is based on further analysis of the Dutch Raven data. Flynn reports that grouping the subjects by educational level makes very little difference - the gains appear almost undiminished in each such group considered individually. Finally, the effects of schooling do not fit the overall pattern of test-score gains; schooling affects tests of content more than tests of reasoning, and the rise in test scores shows exactly the opposite pattern.
Culture, in a general sense, has undergone enormous change in all "modern" countries where test scores have risen. Perhaps the most striking 20th-century change in the human intellectual environment has come from the increase in exposure to many types of visual media. From pictures on the wall to movies to television to video games to computers, each successive generation has been exposed to far richer optical displays than the one before. People once regarded pictures as museum pieces or as occasional decorations for the homes of the rich; now they are everywhere, and everybody takes their own photographs. Schoolchildren of all ages devote far more time to visual "projects" today than they did a generation ago. They devote correspondingly less time to the old "three Rs" of reading, writing and arithmetic, with the predictable consequence that skills in those domains have diminished.

Beyond merely looking at pictures, we analyze them. Picture puzzles, mazes, exploded views and complex montages appear everywhere - on cereal boxes, on McDonald's wrappers, in the instructions for assembling toys and in books intended to help children pass the time. Even the answer sheets for standardized tests - often on pages separate from the questions - assume that the test-takers can locate the right places to record their responses. And static displays such as pictures and diagrams are only the beginning. We have had movies since the 1920s, television since the 1950s and video games since the 1970s. Patricia Greenfield of the University of California at Los Angeles argues that children exposed to these media develop specific skills of visual analysis, skills in which they routinely excel their elders. The assumption that children can program a VCR more effectively than their parents has become a cliche of American society, one that recognizes an important generational shift.

It is possible, then, that exposure to complex visual media has produced genuine increases in a significant form of intelligence. This hypothetical form of intelligence might be called "visual analysis". Tests such as Raven's may show the largest Flynn gains because they measure visual analysis rather directly; tests of learned content may show the smallest gains because they do not measure visual analysis at all.

Although little direct evidence exists for the visual-analysis hypothesis, it does offer the advantage of focusing our attention on the diversity of mental abilities. Flynn's argument that real intelligence cannot have gone up as much as scores on the Raven assumes that there is a "real intelligence" - some unitary quality of mind not unlike Spearman's g. Abandoning that assumption, we may think instead that different forms of intelligence are developed by different kinds of experience. The paradox then disappears - we are indeed very much smarter than our grandparents where visual analysis is concerned, but not with respect to other aspects of intelligence. This is hardly a final answer, but it may be a useful way of thinking about the worldwide rise in test scores.
APPENDIX C: *PROBLEM SOLVING TEXT WITH SIGNALS*

**Notes:**
1. The signals are indicated using the same colour coding as used in CoREAD and are from the end of the study, after participant 40.
2. There are 21 CoREAD screens (title plus 20 paragraphs) with each screen enclosed in a box.
3. The paragraphs are reproduced as they were viewed by the participants, except for the blue highlighting (see 4).
4. The yellow highlighting is used to indicate the words that the author signalled in the text. These signals were not visible to the participants but are used for analysis purposes (see Results – Research Question 2).

---

**PS**

A problem arises when we have a **goal** - a state of affairs that we want to achieve - and it is not immediately apparent how the goal can be attained. Some valuable **clues to the nature of problem solving** can be found in the everyday metaphors we use to talk about it. It is conventional to think of **abstract states such as goals as metaphorical spatial locations**, and **event sequences as metaphorical paths** leading from one state to another. This spatial conception permeates descriptions of problem solving. We speak of "searching for a way to reach a goal," "getting around roadblocks" encountered along the way, finding a "shortcut" solution, "getting lost" in the middle of a solution, "hitting a deadend" and being forced to "backtrack," "approaching the problem from a different angle," and so on.

This conception of **problem solving as search in a metaphorical space**, which underlies our commonsense understanding, has been elaborated to provide a rigorous **theoretical framework** for the **analysis of problem solving**. Although some of the theoretical ideas can be traced back to Gestalt psychologists such as Duncker (1945), the modern formulation of a **general theory of problem solving as search** is due to Newell and Simon (1972). In their problem-space formulation, the **representation of a problem consists of four kinds of elements** - a **description of the initial state** at which problem solving begins; a **description of the goal state** to be reached; a set of **operators or actions** that can be **taken**, which serve to **alter the current state** of the problem; and **path constraints** that impose additional conditions on a **successful path** to solution, beyond simply reaching the goal (for instance, a constraint of finding the solution using the fewest possible steps).
The problem space consists of the set of all states that can potentially be reached by applying the available operators. A solution is a sequence of operators that can transform the initial state into the goal state in accord with the path constraints. A problem-solving method is a procedure for finding a solution. Problem solving is thus viewed as search - methods are used to find a solution path among all the possible paths emanating from the initial state and goal state. Figure 1 provides a graphical illustration of a search space. Each circle represents a possible state of affairs, and the arrows represent possible transitions from one state to another that can be effected by applying operators. A sequence of arrows leading from the initial state to the goal state constitutes a solution path.

The problem-space analysis immediately yields a mathematical result with brutal implications for the feasibility of solving problems. If at each step in the search any of $F$ operators might be applied, and a solution requires applying a sequence of $D$ steps (that is, $D$ is the "depth" of the search), then the number of alternative operator sequences is $F$ to the power $D$. As $F$ and $D$ get even modestly large, $F$ to the power $D$ becomes enormous. A typical game of chess, for example, might involve a total of 60 moves ($D$), with an average of 30 alternative legal moves available ($F$) at each step along the way. The number of alternative paths would thus be $30$ to the power $60$, a number so astronomical that not even the fastest computer can play chess by exploring every possible move sequence. The fact that the size of the search space increases exponentially with the depth of the search is termed combinatorial explosion, a property that makes many problems impossible to solve by exhaustive search of all possible paths.

Humans, with their limited short-term memories, are actually far less capable of "brute-force" search than are computers. For example, human chess players are unable to "look ahead" more than three or four moves. Yet a human grand master can play superlative chess, better than any computer program yet devised. How can this be? The answer is that humans use problem-solving methods that perform heuristic search - rather than attempting the impossible task of examining all possible operator sequences, people consider only a small number of alternatives that seem most likely to yield a solution. Intelligent problem solving, in fact, consists largely of the use of methods for heuristic search. Some heuristic search methods are very general and can be applied to virtually any problem; others are much more specific and depend on detailed knowledge of a particular problem domain. The development of expertise is largely the acquisition of knowledge that restricts the need for extensive search.
The efficacy of heuristic search depends in part on the nature of the problem to be solved. A major distinction is whether the best possible solution is required, or whether any reasonable solution that achieves the goal will suffice. Heuristic methods are seldom much use in solving best solution problems. An example is the notorious "traveling salesman" problem. This problem involves taking the locations of a number of cities (say, ten) and trying to find the shortest possible route that passes through each of the cities exactly once. Due to combinatorial explosion, this problem has an enormous search space of possible routes once the number of cities grows at all large. No one has found a method other than brute force search of all possible routes that guarantees finding the shortest route. However, if the goal is simply to find a route that is reasonably short by some criterion, heuristic search methods may be useful. Human problem solvers are particularly good at what Herbert Simon calls satisficing - finding reasonably good but not necessarily optimal solutions.

Search for a problem solution can proceed in either of two directions - forward from the initial state to the goal state, or backward from the goal state to the initial state. Forward search involves applying operators to the current state to generate a new state; backward search involves finding operators that could produce the current state. In general, it is most efficient to search in whichever direction requires fewest choices at each decision point. For example, if there is only way to reach the goal state, it may be easiest to work backward from the goal.

Newell and Simon suggested a small number of general heuristic search methods. One of the most important of these, means-ends analysis, involves a mixture of forward and backward search. The key underlying means-ends analysis is that search is guided by detection of differences between the current state and the goal state. Specifically, means-ends analysis involves the following steps.

1) Compare the current state to the goal state and identify differences between the two. If there are none, the problem is solved; otherwise, proceed.

2) Select an operator that would reduce one of the differences.

3) If the operator can be applied, do so; if not, set a new subgoal of reaching a state at which the operator could be applied. Means-ends analysis is then applied to this new subgoal until the operator can be applied or the attempt to use it is abandoned.

4) Return to step 1.

Suppose, for example, that you have the goal of trying to paint your living room. The obvious difference between the current state and the goal state is that the room is unpainted. The operator "apply paint" could reduce this difference. However, to apply this operator you need to have paint and a brush. If these are lacking, you now set the subgoal of getting paint and brush. These could be found at a hardware store. So you set the subgoal of getting to a hardware store. And so on, until the conditions for applying the operator are met, and you can finally reduce the difference in the original problem.
Means-ends analysis illustrates several important points about intelligent heuristic search. First, it is explicitly guided by knowledge of the goal. Second, an initial goal can lead to subsequent subgoals that effectively decompose the problem into smaller parts. Third, methods can be applied recursively; that is, in the course of applying a method to achieve a goal, the entire method may be applied to achieve a subgoal. Thus, in step 3 of means-ends analysis, the method may be reapplied to the subgoal of reaching a state in which a desirable operator is applicable. (Recursion is also an important property of human language.)

The idea that the process of problem solving is a kind of search suggests a separation between the initial planning of a solution and its actual execution. It is usually advantageous to perform at least a partial search by "looking ahead" for a solution before actually applying any operators. The obvious advantage of planning is that by anticipating the consequences of possible actions, one can avoid the unfortunate consequences of making overt errors.

The importance of planning varies with the extent to which error recovery is possible. Rich (1983) distinguishes three types of erroneous solution attempts, which she terms ignorable, recoverable, and irrecoverable. An ignorable solution attempt, as term implies, can simply be set aside and another attempt made. For example, if you try to unlock a door with the wrong key, you can simply try another key. The initial error need have no consequences beyond a little wasted time. In contrast, a recoverable solution attempt requires some explicit "undoing" to get back to the state prior to the error. For example, if you find you have made an error in solving a crossword puzzle, you will need to erase the erroneous entry and any other entries that depended on it. Finally, an irrecoverable solution attempt simply cannot be undone. If you are playing chess, for example, you are not allowed to "take back" a regrettable move. Nor can a general who commences a surprise attack on the enemy laterrestore the prior state of affairs. In chess and war as in golf, the problem solver must "play it as it lays."

To the extent that errors are irreversible, or reversible only with difficulty or simply unduly time-consuming, planning is especially important. By imagining the consequences of actions prior to an overt solution attempt one can identify dead ends without actually executing actions. In addition, planning provides information that can be used to monitor and learn from an overt solution attempt. By explicitly anticipating the consequences of applying operators, the problem solver generates expectations that can be compared to what actually happens when the operators are applied. If the actual effects of operators differ from their expected effects this may trigger revision of the plan as well as revision of beliefs about what will happen in similar future applications of the relevant operators. Problem solving thus provides valuable information that can guide learning.
Planning often is combined with a process of **problem decomposition**, in which an overall problem is broken into parts, such that each part can be achieved separately. Suppose, for example, that you need to select a slate of officers to run a club. Rather than trying to select people to fill the entire slate at once, it makes more sense to decompose this goal into several subgoals - selecting a president, a treasurer, and so on. Each of these subgoals defines a problem that can be attacked independently. Finding a solution to each subgoal will require fewer steps than solving the overall compound goal. Because search increases exponentially with the number of steps, solving all the subgoals, each of which requires a relatively small number of steps, is likely to require far less total search than would have been needed to solve the entire problem at once.

Unfortunately, realistic problems are seldom perfectly decomposable into parts that can be solved completely independently. In our example, choices of officers for the various positions interact in various ways. For example, the same person cannot be both president and treasurer and the various officers need to get along with each other - Sally might make a fine president and Joe a good treasurer, but if they dislike each other they would make a poor combination. But despite this lack of complete independence, total search may be minimized by first making some tentative decisions about each subgoal and then later working on integrating the components into a workable overall plan. That is, the problem solver can take advantage of the fact that some problems are **partially decomposable**. This can best be done if foresight is used to form a coherent overall plan before actually beginning an overt solution attempt. Thus, before actually proposing a slate of officers, we could check for compatibility of the tentative list of choices and make corrections where needed. The **general strategy is to first try to solve each subgoal independently, but to note constraints on how the individual decisions interact, and then to check that these constraints are satisfied before finalizing the overall solution attempt.** Planning is thus particularly important in effectively reducing search for partially decomposable problems.

Newell and Simon's **problem-space analysis is highly abstract** and is potentially compatible with a variety of specific representations and algorithms. In practice, however, their approach has been closely tied to a particular type of **formal model**, the **production system**. The central component of a production system is a set of **production rules** (also termed **condition-action rules**). A typical production rule might be

**IF** you have a paint roller
    and you have paint
    and you have a surface ready to paint
    and the surface is large
    and your goal is to paint the surface
**THEN** roll the paint onto the surface
    and expect the surface to be painted.
This rule represents the knowledge required for appropriate application of a problem-solving operator. The "then" portion of the rule specifies the action to be taken and the expected state change it will bring about; the "if" portion consists of a set of clauses describing when the operator could and should be invoked. Note that the clauses in the condition of this rule are of two types. The first four describe preconditions that must be met before the operator can be applied - you need a roller before you can roll. The fifth clause specifies a goal for which the operator is useful - if you want to paint, consider using a roller. The goal restriction helps to limit search, because it means this rule will only be considered when the relevant goal has arisen.

A typical production system operates by cycling through the following steps.

1) The conditions of rules are matched against the currently active portion of memory (for instance, the representation of the current problem state) to identify those rules with conditions that are fully satisfied.

2) If more than one rule is matched, procedures for conflict resolution select one of the matched rules.

3) The selected rule is fired; that is, its action is taken.

4) Return to step 1.

Production-system models of problem solving have been extremely influential in the development of modern cognitive science. Within artificial intelligence such models have been used to develop expert systems that help perform such tasks as medical diagnosis and mineral exploration. In cognitive psychology Anderson's ACT model (1983) is predicated on the claim that human cognition is fundamentally a production system. Alternative models of this general form have many important differences; nonetheless, rule-based systems have provided a common theoretical language that fosters communication among researchers in several of the disciplines that make up cognitive science. Their successful applications have also spread the influence of Newell and Simon's approach to problem solving and to cognition in general.

Production-system models have several virtues. First, they provide a direct method of instantiating knowledge about how to traverse a problem space by applying operators. Second, the knowledge is clearly procedural. Although our example rule looks very linguistic, rules differ from simple descriptions in that they act as instructions for appropriate action. Third, the knowledge is encoded in a highly modular fashion. It is relatively easy to add new rules to the system without unduly disrupting the operation of the older rules. Production systems are therefore capable of modeling learning from problem-solving experience.
APPENDIX D: FIGURE FOR THE PROBLEM SOLVING TEXT

Figure 1
A graphical illustration of a search space for a problem.
APPENDIX E: DEMOGRAPHIC QUESTIONNAIRE - FLYNN EFFECT TEXT

PARTICIPANT - ____________ FE

Demographic Data Questionnaire

Please indicate your response to the following by circling one of the options or with a written answer as appropriate.

Age: ____________

Gender: M F

University Degree Program: BA BSc BEd BEng BComm Other: ____________

Major(s): ____________

Minor(s): ____________

Length of Degree (years): 1 2 3 4 5

Current Year: 1st 2nd 3rd 4th 5th

Number of Psychology courses taken: ____________

Number of courses taken in the Cognitive Sciences other than Psychology: ____________

(e.g. Linguistics, Philosophy of Mind, Artificial Intelligence)

Please indicate your responses to the following by circling "have" or "have not"

I have / have not taken a statistics course.

I have / have not taken a course where a major topic was psychometrics.

I have / have not taken a course where a major topic was intelligence or intelligence testing.

Please indicate your responses to the following by placing an X in the column of your choice

Before having read this text my knowledge of the following topics and concepts was...

<table>
<thead>
<tr>
<th>Topic or Concept</th>
<th>Very Low</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
<th>Very High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Types of intelligence tests</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calculation of IQ scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g (as it relates to IQ)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean of IQ scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard deviation of IQ scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Flynn Effect</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Normal (Gaussian) Distribution</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raven’s Progressive Matrices</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Possible causes of the Flynn Effect</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factor Analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX F: DEMOGRAPHIC QUESTIONNAIRE - PROBLEM SOLVING

Demographic Data Questionnaire

Please indicate your response to the following by circling one of the options or with a written answer as appropriate.

Age: __________
Gender: M F

University Degree Program: BA BSc BEd BEng BComm Other: __________

Major(s): ______________________________________________________________________
Minor(s): ______________________________________________________________________

Length of Degree (years): 1 2 3 4 5
Current Year: 1st 2nd 3rd 4th 5th

Number of Psychology courses taken: __________
Number of courses taken in the Cognitive Sciences other than Psychology: __________
(e.g. Linguistics, Philosophy of Mind, Artificial Intelligence)

Please indicate your responses to the following by circling "have" or "have not"
I have/have not taken a Cognitive Psychology course.
I have/have not taken a course where a major topic was Problem Solving (human or artificial).
I have/have not taken a course where a major topic was Artificial Intelligence (AI).

Please indicate your responses to the following by placing an X in the column of your choice

Before having read this text my knowledge of the following topics and concepts was...

<table>
<thead>
<tr>
<th>Topic or Concept</th>
<th>Very Low - None</th>
<th>Low</th>
<th>Low Moderate</th>
<th>High Moderate</th>
<th>High</th>
<th>Very High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem solving as search</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Problem space metaphor (states, operators, paths)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heuristic search</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Problem decomposition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Means-Ends Analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Production Rules (Condition-Action Rules)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combinatorial Explosion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exhaustive search</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Satisficing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The ACT Production System Model (or ACT*, ACT-R)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX G: CONSENT FORM

PARTICIPANT: ____________________________ Study 2008

Consent Form

Principal Investigator
Andrew Chiarelli
PhD candidate
andrew.chiarelli@mail.mcgill.ca
514-398-4914

Researcher

Faculty Supervisor
Dr. Susanne P. Lajoie
Professor and Chair
susanne.lajoie@mcgill.ca
514-398-3429

McGill University
Faculty of Education
Dept. of Educational & Counselling Psychology
3700 McTavish Street
Montreal, QC, H3A 1Y2

The Study:
This study will examine how reading software that provides information to students based on previous students’ actions leads to the emergence of a stable collective ‘opinion’ or ‘knowledge’ about the text being read. The software that students will use allows each student to read a text and highlight passages they deem important. The software also aggregates these actions and presents future students with information regarding the relevance of each word in the text (high, moderate, low). In this way, the collective decisions of previous students are available as information to be considered by future students.

The Task:
You will be asked to read two texts on a computer using a particular computer program. The software will enable you to highlight the text and will show you what previous students thought were the relevant parts of the text. After you have finished reading each text you will write a summary. I will then ask you to complete a short demographic questionnaire (e.g. degree and major, courses taken related to text topics, etc.). Upon completion I will tell you a little about the rationale for the study and answer any questions you might have. The task should take approximately 2-2.5 hours and you will receive $40 for participating.

Your rights as a participant:
As a participant in a study you have several rights. Your participation is entirely voluntary. Your instructor is not involved in this study and will not know who is participating. You may choose to withdraw from the study at any time (before it commences, during the data collection phase, or after data have been collected) simply by informing me of your decision. If you choose to withdraw I will destroy any data that you provided. Choosing not to participate or withdrawing from the study at a later time will in no way affect your grade in the course.

Anonymity:
There is a participant number at the top right of this consent form. All computerized data files, your summary, and your answers to demographic questionnaire will be tagged with this number. Your name will not appear on any documents other than this consent form. I will store the consent forms in a research lab at McGill. The results of this study will be reported in public forums (journal articles, conference presentations, etc.). Such reports will focus on the aggregate results based on a group of participants, not individuals. However, in the event that an individual participant’s data is used to illustrate a particular finding no identifying information will be provided. As such, your anonymity is assured.

I agree to participate in the study described above. I understand my rights as a participant and have received a copy of this Consent Form for my records.

NAME: ____________________________

SIGNATURE: ____________________________

DATE: __________, 2008

I have received $40 for my participation in a study conducted by Andrew Chiarelli on reading comprehension on the date indicated below.

NAME: ____________________________

SIGNATURE: ____________________________

DATE: __________, 2008
APPENDIX H: AUTHOR’S OVERVIEW FOR FLYNN EFFECT TEXT

Note: These are the first two paragraphs from the original article (Neisser, 1997) and constitute an overview of the article by the author.

Average scores on intelligence tests are rising substantially and consistently, all over the world. These gains have been going on for the better part of a century—essentially ever since tests were invented. The rate of gain on standard broad-spectrum IQ tests amounts to three IQ points per decade, and it is even higher on certain specialized measures. In the Netherlands, for example, all male 18-year-olds take a test of abstract-reasoning ability as part of a military-induction requirement. Because the same test is used every year, it is easy to see the mean score rising, in this case, at about seven points per decade.

The cause of these enormous gains remains unknown. At this point, no one even knows whether they reflect genuine increases in intelligence or just the gradual spread of some specialized knack for taking tests. Greater sophistication about tests surely plays some role in the rise, but there are other possible contributing factors: better nutrition, more schooling, altered child-rearing practices and the technology-driven changes of culture itself. Right now, none of these factors can be ruled out; all of them may be playing some part in the increasing scores. Whatever the causes may be, the sheer size of the gains forces us to reconsider many long-held assumptions about intelligence tests and what they measure.