The Pennsylvania State University

The Graduate School

LUDIC ELICITATION: USING GAMES FOR KNOWLEDGE ELICITATION

A Dissertation in
Information Sciences and Technology
by
Yan Cao

© 2014 Yan Cao

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2014
The dissertation of Yan Cao was reviewed and approved* by the following:

William L. McGill
Assistant Professor of Information Sciences and Technology
Dissertation Advisor, Chair of Committee

David L. Hall
Professor of Information Sciences and Technology

Anna Cinzia Squicciarini
Assistant Professor of Information Sciences and Technology

Tao Yao
Associate Professor of Industrial & Manufacturing Engineering

Peter Forster
Assistant Dean for Online Programs and Professional Education
College of Information Sciences and Technology

*Signatures are on file in the Graduate School.
Abstract

Knowledge elicitation from human beings is important for many fields, such as decision support systems, risk communication, and customer preference studying. Traditional approaches include observations, questionnaires, structured and semi-structured interviews, and group discussions. Many publications have been studying different techniques for a variety of data elicitation tasks as well. However, few of them have considered participants’ user experience in the process. One main drawback of these methods is their time consuming and labor intensive nature, because of which participants often lose their interest and attention quickly in data elicitation activities. Innovated by the success of games with a purpose in many fields such as participatory city exploration and community building, we propose to adopt a game approach for knowledge elicitation tasks.

We have developed two browser-based casual games, LinkIT and SortIT, and have applied them for three knowledge elicitation applications: relation elicitation, rank elicitation, and probability elicitation. The LinkIT game elicits relations between variables/concepts and facilitates the construction of relation network structures such as concept maps and Bayesian networks. The SortIT game presents puzzles in the form of multiple-choice questions. This format supports rank elicitation in a pairwise comparison approach. The second application of this game is probability elicitation by using probability intervals or verbal expressions. By comparing the two games with more traditional methods such as questionnaires, we have established the external validity of the games for the three knowledge elicitation tasks. Further, user experience studies conclude that the games improve user experience by forming the elicitation tasks as a play activity and making the activity more interesting, engaging, exciting, and fun. These findings provide positive support for the applications of GWAP for more knowledge elicitation tasks.
# Table of Contents

List of Figures vii
List of Tables viii
Acknowledgments ix

## Chapter 1

**Introduction** 1

## Chapter 2

**LinkIT and SortIT: Two Casual Games for Knowledge Elicitation** 9

2.1 LinkIT .......................................................... 9
   2.1.1 Game Play .............................................. 10
   2.1.2 Game Bots ............................................. 13
   2.1.3 Winning Condition ................................. 13
   2.1.4 Packaging .............................................. 15
   2.1.5 Counter-Cheating .................................. 16

2.2 SortIT .......................................................... 17
   2.2.1 Game Play .............................................. 19
   2.2.2 Winning Condition ................................. 20
   2.2.3 Packaging .............................................. 20
   2.2.4 Counter-Cheating .................................. 21

## Chapter 3

**Relation Elicitation: LinkIT Elicits Relations** 23

3.1 Introduction ............................................... 23

3.2 Literature Review .......................................... 26
   3.2.1 Risk Perception and Risk Communication .......... 26
   3.2.2 Mental Models Approach in Risk Communication .... 27
   3.2.3 Knowledge Structure and Similarity Ratings ....... 28
3.2.4 Structural Knowledge Elicitation

3.2.5 Research Questions

3.3 Method

3.3.1 Experiment Design

3.3.2 Participants

3.4 Results

3.4.1 Construct Validity

3.4.1.1 Influence diagrams and probabilistic networks

3.4.1.2 Estimation-based significantly different links

3.4.1.3 Difference function

3.4.1.4 Categorization

3.4.1.5 Summary

3.4.2 Productivity

3.4.2.1 Summary

3.4.3 Enjoyment

3.4.3.1 Summary

3.5 Discussion

3.6 Conclusion

3.7 Another Application: Educational Performance

Chapter 4

Rank Elicitation: SortIT Elicits Ranks Data

4.1 Introduction

4.2 SortIT for Rank Elicitation

4.2.1 Create Ranks from Pairwise Comparisons

4.2.1.1 EWR

4.2.1.2 Relative SVD

4.2.2 Ranking Correlation

4.2.2.1 Kendall’s $\tau$

4.2.2.2 Average precision correlation coefficient

4.2.3 Research Questions

4.3 Eliciting Ranks from the General Public

4.3.1 Experiment Design and Participants

4.3.2 Data Quality and Construct Validity

4.3.2.1 Cheating Behavior

4.3.2.2 Ranking Correlation

4.3.3 User Experience

4.4 Eliciting Ranks from Domain Experts

4.4.1 Experiment Design and Participants

4.4.2 Construct Validity
Chapter 5

Probability Elicitation: SortIT Elicits Probabilities 71
5.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 71
5.2 A Game Approach for Probability Elicitation . . . . . . . . . . . . . . 73
5.2.1 Research Questions . . . . . . . . . . . . . . . . . . . . . . . . . 74
5.3 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 74
5.3.1 Experiment Design . . . . . . . . . . . . . . . . . . . . . . . . . . 74
5.3.2 Participants . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 76
5.4 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 79
5.4.1 Productivity . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 79
5.4.2 User Performance . . . . . . . . . . . . . . . . . . . . . . . . . . 79
5.4.3 User Experience . . . . . . . . . . . . . . . . . . . . . . . . . . . 81
5.5 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 82
5.6 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 84

Chapter 6

Future Application: Bayesian Network Elicitation 85
6.1 Bayesian Networks . . . . . . . . . . . . . . . . . . . . . . . . . . . . 85
6.2 Constructing Bayesian Networks . . . . . . . . . . . . . . . . . . . . 87
6.2.1 A Causal Mapping Approach . . . . . . . . . . . . . . . . . . . . 87
6.2.2 Quantitative Probabilities Elicitation . . . . . . . . . . . . . . . . 90
6.2.3 User Experiences . . . . . . . . . . . . . . . . . . . . . . . . . . 93
6.3 A Game Approach for BN Elicitation . . . . . . . . . . . . . . . . . . . 93
6.3.1 LinkIT Elicits BN Structure . . . . . . . . . . . . . . . . . . . . . 93
6.3.2 SortIT Elicits Conditional Probabilities . . . . . . . . . . . . . . 96
6.3.3 The Proposed Game Approach . . . . . . . . . . . . . . . . . . . . 96
6.3.4 NodeIT: Eliciting Nodes from the Game Play? . . . . . . . . . . 97

Bibliography 99
# List of Figures

2.1 An example *LinkIT* puzzle ........................................ 11  
2.2 The *LinkIT* game: time out screen ................................. 12  
2.3 The *LinkIT* game: puzzle feedback screens ....................... 14  
2.4 The *LinkIT* game: summary screen of a game .................... 15  
2.5 The *LinkIT* game: debriefing screen ............................. 16  
2.6 A *SortIT* puzzle example ......................................... 17  
2.7 The *SortIT* game: weapons ....................................... 18  
2.8 The *SortIT* game: summary screen of a game ................... 21  
2.9 The *SortIT* game: instruction screen ............................. 22  

3.1 *LinkIT* elicits relations: the more traditional questionnaire approach .................................. 33  
3.2 *LinkIT* elicits relations: difference function vs. probability threshold .......................... 39  

4.1 *SortIT* elicits ranks: example puzzles ............................ 59  
4.2 *SortIT* elicits ranks: 3-choice puzzles ............................ 64  
4.3 *SortIT* elicits ranks: players’ performance ....................... 66  

5.1 The *SortIT* game: eliciting probabilities .......................... 75  
5.2 *SortIT* elicits probabilities: example probability questions ........................................ 77  
5.3 *SortIT* elicits probabilities: user performance results .......... 80  
5.4 *SortIT* elicits probabilities: user performance with regard to puzzles .................. 80  
5.5 *SortIT* elicits probabilities: user performance with regard to puzzle levels .................. 81  

6.1 A simple Bayesian network example ................................. 86  
6.2 From causal maps to BNs: D-map and I-map ......................... 89  
6.3 From causal maps to BNs: elimination of indirect links ............ 90  
6.4 A structured interview that elicits causal relations .............. 90  
6.5 *LinkIT* elicits BN structure: eliminate indirect relations ........ 95  
6.6 *LinkIT* elicits BN structure: input and output .................. 95  
6.7 *SortIT* elicits conditional probabilities for BNs: input and output .................................. 96  
6.8 How the game system elicits BNs .................................. 97
# List of Tables

3.1 *LinkIT* elicits relations: post-experience survey statements . . . . . . . 35  
3.2 *LinkIT* elicits relations: the categorization method for links . . . . . . 39  
3.3 *LinkIT* elicits relations: the categorization results . . . . . . . . . . . 40  
3.4 *LinkIT* elicits relations: validity results . . . . . . . . . . . . . . . . . . 41  
3.5 *LinkIT* elicits relations: productivity results . . . . . . . . . . . . . . . 42  
3.6 *LinkIT* elicits relations: factor loadings of fun and easiness . . . . . . . 43  
3.7 *LinkIT* elicits relations: factor loadings of impression . . . . . . . . . . 44  
3.8 *LinkIT* elicits relations: user experience results . . . . . . . . . . . . . 45  
4.1 *SortIT* elicits ranks: user experience survey . . . . . . . . . . . . . . . 60  
4.2 *SortIT* elicits ranks: ranking correlation coefficients . . . . . . . . . . 62  
4.3 *SortIT* elicits ranks: user experience survey results . . . . . . . . . . 63  
4.4 *SortIT* elicits ranks: validity results . . . . . . . . . . . . . . . . . . . . 67  
5.1 *SortIT* elicits probabilities: user experience survey . . . . . . . . . . . 78  
5.2 *SortIT* elicits probabilities: user performance results . . . . . . . . . 79  
5.3 *SortIT* elicits probabilities: user experience results . . . . . . . . . . 81  
6.1 The matching between verbal expressions and numerical probabilities . . 92
Acknowledgments

I would like to thank the many individuals and entities that supported and helped me to complete this educational experience. I thank my advisor, Dr. William McGill, for all his precious mentoring and guidance in the five years. I also thank all the committee members, Dr. David Hall, Dr. Anna Squicciarini, Dr. Tao Yao, for their comments and suggestions on the dissertation proposal and along the way. I thank my family for having always supporting me pursuing my interest. I also thank these special friends that I have met at Penn State University, Chu Huang, Zhe Liu, Na Wang, Qiaosha Zou. They played an important role in my life, in both big and small ways. I treasure you in my heart.

To all, I say thank you.
Chapter 1

Introduction

Knowledge elicitation has its formal beginnings in the 1980’s in the context of knowledge engineering and expert systems. With the realization of “knowledge is power” and technological advance, many areas grow a need for expert systems [1]. Questions regarding knowledge elicitation and knowledge acquisition become central to both applied and basic endeavors [2]. How can knowledge be effectively elicited from domain experts? A closer look at the literature reveals that more attention and effort are paid to the representation and conceptualizations of knowledge structure such as networks and schemas [3, 4]. Fewer publications focus on the actual knowledge elicitation process and applicable methods.

Researchers and practitioners have developed many knowledge elicitation methods, many of which are adapted from cognitive methods or methods in other disciplines such as counseling, education, and management [5, 6]. In recent years, knowledge elicitation has surfaced in more areas, including human-computer interaction, training systems, and cognitive engineering [7, 8]. Driven by the need, more work is devoted to developing additional methods.

Many elicitation techniques have been widely used in publications, such as struc-
tured and unstructured interviews, ranking, card sorting, and thinking aloud. One important note is that no single method dominates others and no single method is applicable for all knowledge elicitation tasks. The type of knowledge elicited is an important dimension to distinguish the numerous knowledge elicitation methods. For example, some methods (thinking-aloud and interviews) rely heavily on verbal reports, compared to others (similarity ratings and observations), and are thus more applicable for enumerating concepts and decomposing tasks into steps. Another important note is that for any of the methods, there is no single definitive procedure for the application in practices. A systematic review on empirical studies concerning the effectiveness of elicitation techniques concluded that interviews, especially structured interviews, appear to be the most effective approach [9]. Also, no significant effects on the elicitation of intermediate representations, such as visual hints, is found [9]. Below is a short summary of knowledge elicitation methods, divided into four groups.

**Observations**

Observations are usually the beginning of a knowledge elicitation process and provides an overview of the domain [6]. By observations, researchers and practitioners can generate an initial conceptualization of the task and identify potential constraints or issues to be dealt with during later phases. Depending on the nature of the task, observations can occur in a natural or controlled setting. Advantages of observations include (1) the interference is minimized, (2) observations generates a lot of data. On the other hand, one disadvantage lies on the uncertainty embedded in the interpretation of the elicited data. Many modern technologies could be used to facilitate observations, such as video recorders and video analyzing software.
**Interviews**

Interviews provide a direct approach to elicit knowledge by asking the corresponding people (experts). Two forms of interviews, unstructured and structured, are widely used in practice. Unstructured interviews are usually adopted for early stages of elicitation and help the elicitors understand the domain and prepare more structured interviews. Structured interviews have predetermined content and sequencing and thus provide more constraints on the experts’ responses and thus more systematic converge of the domain. Interviews are relatively easy to administer compared to other knowledge elicitation methods. However, they pose a higher pressure on the data analysis and interpretation phase. In recent years, some domains have developed highly specific interview methodologies, such as the mental models approach for risk communication [10].

**Process Tracing**

Process tracing is most often used to elicit procedural information such as sequential behavioral events, including both verbal events and non-verbal events. When and how the data is collected are two important elements for determining the quality of collected information.

**Conceptual Methods**

Conceptual methods are used to elicit concepts and their inter-relations in domains. These methods usually associate with several steps, including (1) the elicitation of domain-related concepts through interviews or document analysis, (2) the determination of relationships between concepts, and (3) the representation and interpretation of the relationships [6]. A number of relationship judgment methods have been developed, such as pairwise similarity ratings, sorting techniques, repertory grid, and frequency of
co-occurrence. The similarity ratings approach involves presenting pairs of concepts to
the expert and requesting a quantitative estimate of the similarity of the two concepts.
It exists a scaling issue and becomes very time costly when the number of concepts
exceeds 30. In the repertory grid approach, all concepts are rated against a set of dimen-
sions and the similarity between concepts can thus be derived from the ratings.

Although many publications have been devoted to knowledge elicitation methods
and applications, few of them have discussed participants’ user experience in this pro-
cess and have not proposed any method to improve their experiences. Our previous
studies in the elicitation of lay people’s risk mental models indicate that interviews and
surveys with human subjects for structural model elicitation purposes are usually very
time-consuming and repetitive in terms of the nature of the tasks. As a consequence,
human participants get bored quite easily and are unable to maintain their interest and
attention. This implies the necessity to study and improve user experience in the process
of knowledge elicitation tasks.

Simulation and gaming scholar Richard Duke argued in his 1974 text that “gaming
is the future’s language” [11]. The sentiment underlying this seemingly radical state-
ment for the time was based on the author’s observations that game mechanics can be
used to stimulate meaningful “multilogue” among players that other approaches would
be hard-pressed to achieve. It is now apparent that after four decades of gaming re-
search, modern gaming has exceeded Professor Duke’s original expectations. Gaming
has become more than a language; it has become an approach to stimulate productivity
of all sorts. Games are routinely used these days to complement classroom instruction
[12], facilitate group discussions about disaster preparedness [13], and incentivize par-
ticipation in data collection and human computation campaigns [14], to name just a few
applications. In fact, contemporary discourse on the future of gaming centers on the
parallel notions of ubiquitous and pervasive gaming, where some philosophers suggest
that gaming has the potential to change the world [15, 16].

Among the more popular forms of contemporary gaming is the “casual game”. A casual game is game distinguished by simple rules and lack of commitment required to become proficient and competitive [17]. Many traditional board games fall under the category of casual game, including such popular titles as CONNECT FOUR, SCRABBLE and SORRY! as well as classic games such as BACKGAMMON, PACHISI, and MANCALA. The Nintendo Wii, a leading video game console, attributes its success through the appeal of casual gaming to non-traditional players [18]. The Internet offers myriad casual games that are free-of-charge to play and take advantage of standard web browsers and associated plugins or extensions (e.g., Adobe Flash) to create a widely accessible and lightweight gaming environment [19]. Such games are known as browser games.

Human computation is a field that studies the employment of human’s computing ability to solve complex problems that computers cannot easily solve in time. Games present one platform for human computation in an entertaining manner. Pioneered by Luis von Ahn and his colleagues, many games with a purpose (GWAP) systems have been developed in recent years, such as ESP [20], GWAP for the Semantic Web [21], and CityExplorer [22]. Time-intensive or otherwise tedious tasks are “crowd-sourced” to the players and solved by human power. Thus, the “serious purpose” of a GWAP, that is, the designer’s intent for creating the game [23], is to incentivize productive work via the allure of fun. Numerous publications have shown the successes of these systems in fulfilling the serious purpose while providing an enjoyable environment for participants [20, 21, 22]. This dissertation focuses on the application of GWAP on several knowledge elicitation tasks including relation elicitation, rank elicitation, and probability elicitation.

Motivated by the concepts of serious games [24] and Games With A Purpose [25, 20, 21, 22], a gaming approach has the potential to improve user experience by making
the activities more entertaining. The basic idea is to form the tasks as a play activity and to deploy players’ knowledge and computing power in their game play. For example, players provide high-quality image tags while playing the ESP game [20]. GWAP has demonstrated its usefulness for many serious purposes, such as building the semantic Web [21] and creating domain-specific sentiment lexicons [26].

The purpose of this research is to propose the application of GWAP for several different knowledge elicitation tasks, including relation elicitation, rank elicitation, and probability elicitation. This gaming system consists of two browser-based online casual games, LinkIT and SortIT. The LinkIT game is a relation game and game players draw the relations between variables to proceed in the game play. The SortIT game presents puzzles in the form of multiple-choice questions and is adaptable for pairwise comparison applications, which is a frequently used approach for rank elicitation. Meanwhile, utilizing the flexibility of the puzzle question prompt, we can use the game for probability elicitation. Combining these applications provide a potential platform for Bayesian network elicitation as well.

However, it is not guaranteed that the data collected in a game approach is as valid and as accurate as the data collected from a “more serious” environment (such as questionnaires and interviews). Thus, our study focuses on the external validity of the game approach for the three data elicitation tasks. In particular, the main research question is as follows.

**Can a game based approach match the effectiveness of knowledge elicitation using more traditional questionnaire approach? In particular, does the data collected from the game approach maintain the validity?**

Further, because the main reason that we introduce the game approach is to improve user experience in the traditionally tedious knowledge elicitation tasks, our second research
question is

Can a game be designed to improve the user experience of knowledge elicitation compared to the more traditional questionnaire approach?

In terms of the three application fields, our research questions are as follows.

Can a game based approach be used to elicit probabilities in addition to Bayesian network structure?

How well does a game based approach elicit probability knowledge, compared to more traditional questionnaires?

We conduct experimental studies for the three knowledge elicitation tasks to explore the external validity and improvement on user experience of the two games. Significant similarity between the data collected from the game approach and the data from more traditional approaches is identified, supporting the external validity of the game approach. Also, participants have rated the game play activity to be more enjoyable and maintain their attention longer than traditional questionnaire activity does.

The contributions of this research are three-folds. First, it highlights the importance of studying user experience in knowledge elicitation processes. Second, we propose a game approach for several different knowledge elicitation tasks and call it “ludic elicitation”. We have developed two casual games for these applications, LinkIT and SortIT. Third, experimental studies are conducted to demonstrate the external validity of the game approach and their improvement on participants’ user experience.

The rest of this dissertation is organized as follows. Chapter 2 introduces the two games that have been developed, LinkIT and SortIT. The following three chapters summarizes our experimental studies on the three knowledge elicitation tasks: relation elicitation, rank elicitation, and probability elicitation, respectively. Each chapter starts
with an introduction to the knowledge elicitation task, followed by the corresponding research question(s). We then describe the experiment design and participants in the method section. The results are presented and discussed afterwards. Each chapter also has a conclusion of the findings. In the last chapter, we further propose a game approach for eliciting Bayesian networks, on the basis of the three completed studies.
Chapter 2

LinkIT and SortIT: Two Casual Games for Knowledge Elicitation

2.1 LinkIT

*LinkIT* is a game of relationships. It is a casual online browser game that presents players with two factors that might or might not be related to each other in a specified context of contemporary interest. The aim of *LinkIT* is to provide the player with an entertaining experience; however, from our point of view, the serious purpose of *LinkIT* is not to cash in on ad placement or subscription fees, but to collect data on player perceptions toward the relations between the factors. That is, each round of play generates data pertaining to how the player views relationships in the context.

*LinkIT* was written in Flash by a team of graduate and undergraduate students at the Pennsylvania State University, and is publicly accessible online at [http://riskgames.ist.psu.edu](http://riskgames.ist.psu.edu). A socket connection between the server and the client supports game play, in which the server provides the client with puzzles and feedback while the client sends back players’ responses to puzzles. The server also stores game play data in a
secure relational database.

Two versions of LinkIT have been designed. In the “duo” version, players are paired together through a random selection on the server. Players do not have to be collocated to be paired; they can be anywhere in the world with an Internet connection, but they must be logged in at the same time. The paired players collaborate with each other to make progress in the game. The identity of a collaborator is not revealed, which stifles communication between partners outside of the game; this strategy is used to partially mitigate the risk of cheating. In the “solo” version, players are paired with computer agents (Section 2.1.2) and are informed that they are playing with another player. Thus, the two versions are both 2-player games and are essentially similar to each other, the difference being who the other player is, whether human or machine. In other words, the player experience should be the same in both cases.

2.1.1 Game Play

LinkIT is a puzzle game and the puzzles are composed of several nodes, one relation word, and a context. The nodes are concepts, events, or circumstances within a given context. For example, in the climate change context, example nodes could be “the use of CFCs”, “CO2 emission”, and “Industrial solvents and lubricants”. The relation word, for example “causes” or “can lead to”, specifies the type of relationship between nodes that the game explores. Short descriptions are attached to nodes and appear on the screen when the mouse cursor hovers over nodes. Figure 2.1 provides an example puzzle in the climate change context, containing two nodes, “CH4 Emissions” and “Industrial Solvents and Lubricants”, and the relation word “Causes”. Players respond to the puzzle by drawing the directional relationship they think exists between the two nodes, such as “Industrial Solvents and Lubricants Causes CH4 Emissions” by clicking on the starting
node and then the ending node. To assist the player in understanding the nodes, a short description of what is meant by each can be obtained by hovering over the node with the mouse cursor.

![Image of a LinkIT puzzle](image)

**Figure 2.1.** An example of a *LinkIT* puzzle containing two risk events represented as nodes and one relation word.

The puzzles presented to the players are selected from a pre-determined set of nodes and relationship words, from a previously developed expert model. For example, under a climate change context, the game can explore whether there is a perceived causal relationship between “CO2 Emission” and “Global Warming”. For different context and different studying purposes, the game supports a lot other relation words, including “May Cause”, “May Affect”, “In the form of”, “Attributed to”, “Protects”, “Requires”, and so forth. Changing the relation word is simply a matter of adding an entry to the puzzle database.

For each puzzle with two risk nodes A and B and a relation word C such as “May Affect”, players have several possible responses. They can draw any number of non-repeated links between nodes or click the “No Link” button at the center of the screen to indicate the independence between the nodes. Also, they can skip puzzles without pro-
viding an answer if they are unsure about the link. In summary, the 5 possible responses to a given puzzle (A, B, “May Affect”) are: (1) A may affect B, (2) B may affect A, (3) A may affect B and B may affect A, (4) no relation exists between A and B, and (5) I don’t know/Skip this puzzle.

To create a challenge, LinkIT places a time limit on each puzzle. As time progresses in a particular puzzle, the nodes become increasingly agitated as indicated by the intensity in which they appear to vibrate. When both players submit their response to a puzzle, the system evaluates whether they agree and award extra play time on this basis. If players neither click the “Done” button nor the “Skip” button within a certain amount of time, the nodes will “pop” and splatter “paint” all over the screen (see Figure 2.2). Either case marks the end of a puzzle. So long as time remains on the clock, the game advances to the next set of nodes.

![Image of LinkIT game](image)

**Figure 2.2.** The LinkIT game: when too much time passes before both players submit their responses to a puzzle, a timeout screen shows up.
2.1.2 Game Bots

In the solo version, players are paired with computer agents to make progress in the game. We have developed a few different types of these game bots. The first one is an expert bot who plays the game according to “experts” opinions represented internally as an “expert” influence diagram. For each puzzle, the bot draws links that represent what experts perceive as the relationship between the given risk events. In this way, we can potentially extend the game as an educational tool and are presently studying this application.

This expert model can be created only when there exists an expert mental model in the given context. When such a model does not exist, LinkIT employs an adaptive bot that plays the game according to all previous puzzle data; that is, the adaptive bot plays the consensus answer among all previous players presented with the same puzzle. When insufficient response data exists to form a consensus response, the bot either plays as an expert or provides a random answer. The third bot plays the game randomly. For each possible relation (such as “A May Affect B”), the random bot has a 50% probability to draw the relation and the other 50% probability to not draw it.

2.1.3 Winning Condition

LinkIT is an output-agreement game [27] similar to some other GWAP systems such as ESP [20] and Matchin [28]. Players are rewarded only when they produce the same answers as their partners. Because a player cannot see the partner’s response or communicate with the partner during the game play, the easiest way for them to produce the same output is by providing something related to the common output. With regard to risk perception, we assume a common perception exists in a group of lay people. If this assumption is true and players agree with this assumption, players will provide what
they personally believe the answer is in order to maximize the chances of winning. As noted in the instructions provided to players at the start of the game, players are also encouraged to provide their true perceptions for the best possibility of winning the game. Thus, this output-agreement game format acts as a facilitator for better elicitation. Figure 2.3 shows a pair of feedback screens, one when the players agree on a response to the puzzle (win), the other when the players disagree (fail).

Figure 2.3. The LinkIT game: Pair of feedback screens for Match (win) and No Match (fail) of one puzzle.

Players’ achievements are reflected as the maximum time-on-game. That is, LinkIT is a survival-based game, a game where accomplishment is reflected in how long a player (or pair of players, as in this case) can keep playing before losing, such as by running out of time. Each game starts with a countdown clock, the end of which indicates a game end. The reward for a successful match between players is to add time to the clock, whereas the punishment for an unsuccessful match is to take time off the clock. Successive matches amplify the reward (e.g., two successive matches awards 10 seconds, three 15 seconds, and so on), whereas successive fails amplify the punishment. At the conclusion of the game, the player’s performance is summarized on screen as shown in Figure 2.4.
Figure 2.4. The *LinkIT* game: Summary screen presented to the player after time runs out and the game is over.

### 2.1.4 Packaging

We built both priming and debriefing, collectively referred to as “packaging”, into *LinkIT* to help players better understand the game and think about how their opponents responded to the different puzzles. When players first enter the game screen an instruction page is displayed that presents players with the purpose of the game and explains the game play. After a game is over, players have the option to “View Results” – the debriefing phase. In simulation games, debriefing is to invite players to make a connection between experiences in the game play and experiences in real-life situations and thus to encourage players learn from the game [29]. At the debriefing phase in *LinkIT*, players can view their responses and their partner’s responses to all puzzles, such as is shown in Figure 2.5. In this way, players learn others’ understanding of risk. If either player wishes to comment on this link in any way, they have the option to by clicking on “Discuss” and submitting feedback.
Figure 2.5. The LinkIT game: A LinkIT debriefing screen that displays the responses of a player as well as the responses of his/her partner to the each puzzle. Players can enter their comments by hitting the Discuss button.

2.1.5 Counter-Cheating

Previous studies conclude that players tend to adopt a cheating strategy by submitting consecutive No Links to puzzles [30]. In the current version of the game, we have implemented algorithms to detect this cheating behavior for better data quality. Once a player submitted No Link for 8 puzzles in a row (which coincides with a probability of cheating of near 100%), the system ends the game and removes the game play data for these 8 puzzles. Players are displayed with a statement: “Either you or your partner might be engaged with some sort of cheating behavior and we need to end this game. Sorry for that. You can click Play Again to start a new game”. The implementation of game bots also helps us counter cheating, because cheating only works when the two players paired together for one game both cheat. Paired with a bot, who always play “honestly”, a human player is forced to play honestly as well or else the player will quickly lose.
2.2 SortIT

The SortIT game is a game of choices – it presents puzzles in the form of multiple-choice questions. This format supports applications of the game for different contexts, such as ranks data elicitation and probability elicitation, which will be described in Chapter 4 and 5. This section introduces the game mechanics.

SortIT is an online casual game with a similar format as the LinkIT game, in which players make choices instead of making links between nodes. Because our focus is how the gaming platform re-formats knowledge elicitation tasks, rather than creating a fancy game product, the SortIT game we have developed is not implemented for usage at a large scale. It has a simple story background and challenges players by allocating time pressure in game puzzles. Figure 2.6 shows an example puzzle in the SortIT game.

![SortIT puzzle](image)

**Figure 2.6.** An example of a SortIT puzzle containing two risk events represented as nodes and one question “which one presents more risk?”. The context is Emergency Medical Dispatch categories.

SortIT was also written in Flash by the same research team at the Pennsylvania
State University and is publicly accessible online at http://riskgames.ist.psu.edu. A socket connection between the server and the client supports game play, in which the server provides the client with puzzles and feedback while the client sends back players’ responses to puzzles. The server also stores game play data in a secure relational database. The same as LinkIT, we have two versions of the game: a “duo” version and a “solo” version. In the “solo” version, several robot players are designed to either play as an expert, or play as an average player.

One aspect that makes SortIT different from LinkIT is that SortIT has a very simple “story” background in the game play. Players fight with monsters (tigers, dinosaurs, and lions, etc.) and need to defeat them to proceed in the game play. In addition, players earn and lose weapons in the game play, which increase players’ ability of either defending or attacking. Figure 2.7 shows the screen in which players equip weapons before playing a game. Literature indicates that a story element makes the game more enjoyable [31].

**Figure 2.7.** In SortIT, players earn weapons and can equip them before playing a game.
2.2.1 Game Play

As shown in Figure 2.6, a SortIT puzzle is composed of one given context and several nodes, from which players make choices. In general, players make comparisons between events (represented as nodes) in terms of the given context. In this example, players are asked to choose, between remote desktop and unprotected networks, the one that presents more risk.

For each puzzle with several nodes, players make one and only one choice by clicking the corresponding node. In order to assist players in understanding the nodes, a short description of what is meant by each can be obtained by hovering over the node with the mouse cursor. Each puzzle is selected from a pre-determined set of nodes and context expressions, from a previously developed model. For example, in the example shown in Figure 2.6, the context expression is “Which presents more risk?” and the two nodes are selected from a pool of risks in the cyber-security field. The context expressions could be any other things such as “Which of the following better describes brainstorming?”.

Similar to the LinkIT game, SortIT places a time limit on each puzzle to create a challenge for players. As time progresses in a particular puzzle, players get warning sounds. When both players submit their response to a puzzle, the system evaluates whether the responses are the same and increase/decrease players’ “health level” (named as “Health Bar” on the screen) on this basis. If one or both players made no clicking action within a certain amount of time, this puzzle ends. Players have the option to “Skip” a puzzle with no penalty on the “health level”. As long as both players’ “health level” remains positive, the game advances to the next puzzle.
2.2.2 Winning Condition

SortIT is also an output-agreement game [27]. Players are rewarded only when they produce the same answers as their paired partners. Because a player cannot see the partner’s response or communicate with the partner during the game play, the easiest way for them to produce the same output is by providing something related to the common output. We assume that a common sense exists with regards to all puzzles provided in the game. This is not to say that all players have exactly the same perceptions toward the puzzles. Individuals perceive things in their own ways. However, we assume that the general public would reach an agreement on the puzzles. As noted in the instructions provided to players at the start of the game, players are also encouraged to provide their true perceptions for the best possibility of winning the game. Thus, this output-agreement game format acts as a facilitator for better elicitation.

With regards to a winning condition in the game play, interestingly, players can never win the game – a player will keep updating his/her health level and weapon list while making progress in the game play. New monster comes up if a monster is beaten to death. If a player’s health level reaches zero, the game ends and the player looses. Players’ achievements are reflected as weapons earned or lost, as shown in Figure 2.8.

2.2.3 Packaging

We built both priming and debriefing, collectively referred to as “packaging”, into SortIT to help players better understand the game and think about how their opponents/collaborators responded to the different puzzles. When players first enter the game screen, an introduction page is displayed where players can choose from three types of instructions to read through for understanding the gameplay. Figure 2.9 is a screenshot from one of the instructions, displaying all elements on the game play screen. Readers
At the end of each game, players review the weapons he/she has earned or lost during the game play. Also, all puzzles appeared in this game are summarized in a table and players can learn his/her partner’s response to each puzzle.

After a game is over, players have the option to “View Results” – the debriefing phase. As introduced in Section 2.1.4, debriefing is to invite players to make a connection between experiences in the game play and experiences in real life situations. Thus, debriefing is to encourage players learn from the game [29]. Figure 2.8 shows the debriefing screen in the SortIT game, from which players can review the game puzzles and learn his/her partner’s response to each puzzle.

### 2.2.4 Counter-Cheating

Informal studies with students playing the game conclude that players might make consecutive left (or right) clicks to see whether the partner will adopt the same strategy. However, the positioning of nodes is random to any player and thus the two paired play-
Figure 2.9. A screenshot from one of SortIT instructions, displaying and explaining all elements on the game play screen.

Players might not get the same sequence of nodes. This cheating strategy therefore does not work. After a few tries, most players will abandon this cheating strategy.
Chapter 3

Relation Elicitation: LinkIT Elicits Relations

The LinkIT game is designed to elicit relations, i.e., structural knowledge. This chapter presents an experimental study that explores the construct validity and performance of LinkIT as a network structure elicitation tool. Readers should note that this chapter is adapted from a published work [32], which focuses on the application of relation elicitation for risk communication.

3.1 Introduction

The field of risk communication is always trying to find new and creative ways to provide laypeople with information to engage in meaningful discourse about risk and make better and more informed decisions whose outcomes are clouded by risk and uncertainty. As suggested by James Matschulat, “the keystone of risk management is understanding” [33]. The key work of Morgan, Fischhoff, Bostrom, & Atman (2001) notes that risk communication is “intended to supply laypeople with the information
they need to make informed, independent judgments about risks to health, safety, and the environment” [10]. This statement echoes Matschulat who emphasized understanding as a prerequisite to effective decision making to manage risk. Risk communication should only communicate essential information while not omitting any critical information. Naturally, to do this effectively requires us to establish what experts know about risk, understand what laypeople presently believe about risk, identify differences in understanding between experts and laypersons, and then design communication materials that seek to narrow any identified knowledge gaps and correct misperceptions. Thus, the process of developing a risk communication can be distilled into efforts aimed at addressing the following three general questions of risk communication:

To accomplish the difficult task of generating a risk communication, Morgan et al. (2001) established the “mental models approach” to risk communication [10]. The mental models approach begins with the construction of an expert mental model about a particular risk expressed as an influence diagram. Concurrent with the development of expert mental models, layperson models are constructed from data collected from focused groups, one-on-one interviews and surveys [34, 35, 36]. With models representing both expert and layperson understanding of risk in hand, risk communication focuses on resolving the discrepancies between expert and layperson mental models of risk. Communication materials are drafted, evaluated and ultimately disseminated to correct laypeoples misconceptions and ignorance about the nature of risk and factors that influence the chance, extent and magnitude of potential harm.

A significant challenge associated with implementing the mental models approach is the expense associated with constructing both expert and layperson mental models for different societal risks. The present scheme of engaging experts and laypeople face-to-face and one-at-a-time presents a significant time and financial cost. Depending on the stakes of the risk communication context and purpose, sometimes even tens of in-
terviews is not enough. This challenge is further aggravated by the fact that researchers also desire to understand how different segments of the population understand risk and how this understanding evolves over time and in response to different world events [37]. Fortunately, there is a contemporary solution to many of these drawbacks that could potentially address the tediousness of some aspects of the mental models approach. Inspired by several decades worth of research from the simulation and gaming community, recent advances in Games with a Purpose, or GWAP technology, and the practical need to improve the means in which the risk community captures layperson data for risk studies, we propose the concept of ludic elicitation for eliciting data in support of risk studies. By our definition, ludic elicitation is a process of eliciting information through playful activities, such as via an online game. The basic idea is that, as scholars recently noted, a lot can be achieved when an otherwise tedious task is encapsulated in an entertaining activity.

The concept of Games with a Purpose (GWAP) [25], or the leveraging of society’s willingness and desire to play games in order to encourage productive work, suggests an innovative method to advance the mental models approach in terms of making it more efficient and fun. Guided by this thought, we have designed the LinkIT game to facilitate the elicitation of lay people’s perception about the relationships between risks [30]. By virtue of simply playing LinkIT, players communicate their understanding of risk through the choices they make in the game. This game is our first attempt to make use of a GWAP for studying public perceptions of risk. Consequently, a study to establish the games external validity and examine its effectiveness is a necessary prerequisite to leveraging the data collected by LinkIT to inform the creation of risk communications.

In this study, we utilize a between-subject experiment design to study the external validity of LinkIT as a ludic elicitation tool to assist the mental models approach. The
remainder of this chapter is arranged as follows. We first briefly review the concept of risk communication and the mental models approach, thus establishing the need for the \textit{LinkIT} game. After illustrating the \textit{LinkIT} game mechanics, we propose three research hypotheses toward the games external validity, productivity, and enjoyment. We then describe the method in the experiment, and present data analysis results. We further discuss the findings, limitations, and future directions. The last section concludes the article.

3.2 Literature Review

3.2.1 Risk Perception and Risk Communication

Risk communication, in general, is the flow of risk information and evaluations between experts, interest groups, practitioners, and the public [38]. Most literature focuses on the delivery of information from experts to lay people, though information may also flow between experts and from experts to decision makers. Five goals are frequently pursued in the communication process, which are building trust, creating awareness, enhancing understanding, developing agreement, and motivating actions [39, 40]. Risk communication characteristics, such as intentionality, content, audience for whom the communication is directed, source of information, and flow of message, affect the lay people’s risk perception, pertaining to such things as the severity of risks and acceptable actions to control risk [41].

The opposite direction of information flow, i.e., experts trying to understand the publics risk perception, is also indispensable. Risk perception is a determinant factor that influences peoples decisions and responses to risks and risk assessment activities [42]. Risk communication and risk assessment should take account of public concerns
and attitudes, which directly impact human health and social regulations [43, 44]. Understanding lay peoples risk perception helps experts (1) identify common misconceptions and knowledge gaps in the public, and (2) discover distrust and reasons for it. Accordingly, experts can prioritize risk communication contents and methods [42]. The LinkIT game is designed to elicit the lay peoples risk perception, in terms of the relationships between risks.

3.2.2 Mental Models Approach in Risk Communication

Atman, Bostrom, Fischhoff, and Morgan proposed the mental models approach to guide the design and evaluation of risk communications [42, 34]. Influence diagrams are used to represent experts and lay peoples mental models. In such a directed network, nodes are events in the probabilistic process and links describe the directed (e.g., causal) relations between risks. The adoption of influence diagrams is not innovative here, which have been used in other fields such as decision analysis to represent knowledge maps. Many researchers employed this approach and demonstrate its capability in identifying and prioritizing risk communication contents under different contexts [45, 46]. Some publications demonstrated the efficacy of this approach through multiple strategies, such as mailed questionnaires, semi-structured verbal protocols, and user discussion groups [47].

Morgan and collaborators provided a systematic description for the mental models approach for risk communication, which is comprise of four steps [10]. First, an influence diagram representing expert opinions with regard to the subject matter is created. It determines the scope and guides the following open-ended interviews, which are conducted on lay people (step 2). Careful attention should be paid to exclude the possibility of either inducing or dispelling misconceptions. As a result, this step is usually time and
labor intensive and thereby limits the number of interviews that can be conducted. After identifying potential common perceptions and misunderstandings among the public from the interviews, researchers or experts can distribute confirmatory questionnaires to the public, focusing on these identified perceptions (step 3). The final step is the development and evaluations of risk communication.

### 3.2.3 Knowledge Structure and Similarity Ratings

The term “mental models” can be assimilated as a term discussed in fields such as communication and knowledge management, knowledge structure, which describes peoples’ perceptions of the relationships or interconnections between concepts or events. Jonassen, Beissner, and Yacci offer a broad review on techniques for representing, conveying, and acquiring structural knowledge [48].

The similarity ratings approach provides a simple and direct tool for knowledge structure elicitation and representation. It adopts a spatial metaphor in which concepts with more closely related meanings are closer to each other than semantically dissimilar ones [48]. In a typical similarity ratings approach, users are presented with pairs of items and asked to rate the similarity between them, a symmetric undirected relation type. The items are selected from a reference expert model. This approach can be easily extended to acquire perceptions of asymmetric directed relations such as causal relations between risks.

Thus, a potential application is to use the similarity ratings approach to decrease the number of open-ended interviews and replace the confirmatory questionnaires in the mental models approach. In this modified version, after constructing an expert model, we could conduct a few interviews with lay people with an emphasis on their terminology on the subject matter. The data provide a set of nodes from which we draw pairs and
use the similarity ratings approach to let lay people describe the relations. In this way, we may not gather enough data such as common misconceptions from the limited number of surveys. Nonetheless, the similarity ratings allow us to display any pair of nodes, retrieve the publics perception on any relations, and hence eliminate the possibility of missing information.

Although this relieves the intensive time and labor requirement of the interviews, the scaling issue of the similarity ratings approach introduces another challenge. Let \( V \) denote the number of nodes to be covered. The number of pairs to be investigated is thus \( V(V-1)/2 \) for symmetric relations and \( V(V-1) \) for asymmetric ones. When \( V \) is large, it is not feasible to display all pairs to one person. For example, 25 nodes will require 300 or 600 similarity ratings for a comprehensive coverage.

In organizational science, the notion of group cognition and group mind is sometimes adopted to study team performance [49]. Terms such as group mental models, shared mental models, or team mental models are frequently used. The concept is consistent with the purpose of risk communication. Experts are more interested in how the general public perceives risks and risk relationships and aim to make decisions with society-level benefits. This aspect eases the scaling issue because not all pairs need to be answered by each individual. However, because of the limited format and repetitions in the similarity ratings process, users may lose their interest and attention very quickly. This concern highlights the need for a more interactive and enjoyable environment.

In what follows, we describe the LinkIT game designed to elicit players risk perceptions as expressed in terms of the relations between risks. The puzzle interface is similar to the similarity ratings approach discussed previously, in the sense that each puzzle examines one pair of nodes (Section 3.1), but presented in a gaming environment.
3.2.4 Structural Knowledge Elicitation

The task of causal structure discovery is a fundamental problem in many applications, such as the construction of Bayesian networks (BN) and building causal maps [50]. A knowledge structure, describes an individual’s or group’s perceptions of the relationships or interconnections between concepts or events. Jonassen, Beissner, and Yacci offer a broad review on techniques for representing, conveying, and acquiring structural knowledge [48].

Besides open-ended interviews and confirmatory surveys, the similarity ratings approach provides a simple and direct tool for knowledge structure elicitation and representation. It adopts a spatial metaphor in which concepts with more closely related meanings are closer to each other than semantically dissimilar ones [48]. In a typical similarity ratings approach, users are presented with pairs of items and asked to rate the similarity between them, a symmetric undirected relation type. The items are selected from a reference expert model. This approach can be easily extended to acquire perceptions of asymmetric directed relations such as causal relations between risks.

Compared to the interviews approach, the similarity ratings or the pairwise comparison approach relieves the intensive time and labor requirements. However, it also presents a scaling issue and brings new challenges. Let $V$ denote the number of variables to be explored. The number of pairs to be investigated is thus $V(V - 1)/2$ for symmetric relations and $V(V - 1)$ for asymmetric ones. When $V$ is large, it is not feasible to display all pairs to one person. For example, 25 nodes will require 300 or 600 similarity ratings for a comprehensive coverage. The gaming system I propose in this research will utilize a similar format as the similarity rating approach but will ease the scaling issue.
3.2.5 Research Questions

The primary purpose of LinkIT is to elicit game players’ risk perceptions or influence diagrams in general. An essential question to ask is whether the data elicited from LinkIT is valid and accurate. For example, a player may be very eager to win the game and is actually guessing his/her game partner’s perception instead of providing his/her own perception. Admitting the possibility of this and other cases, we propose an hypothesis exploring the construct validity of LinkIT:

H1. Given a completely randomized experiment design and a pre-defined set of risk events as well as a given relation word, the group risk mental model elicited from the LinkIT game is similar to the one elicited from a more traditional approach.

The underlying assumption is that an aggregated group-level mental model exists for all participants, although individuals may have differences in their perceptions. We use a probabilistic network to represent this group mental model, in which each node denotes a risk event and each directed link $A \rightarrow B$ indicates that a causal or influential relation exists from $A$ to $B$. Each link is also associated with a non-negative number between 0 and 1, expressing the percentage of agreement on this relation among the player population. For some links, this number might be very close to 1, indicating that most people agree with this link. If it is close to 0, the majority of the player population does not agree. When the number is somewhat close to 0.5, we say that the player population has a mixed understanding of the particular relation.

Assuming construct validity holds, we expect the game to outperform traditional approaches in two aspects. First is the productivity or the efficiency of the game as an elicitation tool. As mentioned above, pairwise comparisons or similarity ratings are usually adopted to elicit structural knowledge and the number of comparisons is pro-
portional to $V^2$, in which $V$ is the number of elements (items) in the structure. This approach is thus time-consuming and the task soon becomes tedious to participants. Consequently, we are encouraged to examine whether the LinkIT game can make the task less time-consuming, i.e., faster, without losing accuracy.

H2. Under the same amount of time, users playing LinkIT provide more or at least as much data as those responding to a traditional approach.

The second aspect is to tackle the tedious nature of the task. The idea of “ludic elicitation” is change a tedious task to a play activity and thus make the elicitation process fun and enjoyable to the point where players forget they are actually doing work.

H3. LinkIT users report enjoying the activity more so than users responding to a traditional approach.

3.3 Method

We compare the LinkIT game with a more conventional questionnaire-based approach through an experimental study using Amazon Mechanical Turk (MTurk) as the experiment platform and recruitment pool. The reasons we select MTurk are as follows [51]. First, it provides a participant pool arguably closer to the US population as a whole than traditional university subject pools. Second, the low average hourly payment that people are willing to work on MTurk allows us to recruit enough participants for this study given limited funding. On average, our experiment paid participants $4.30 for each hour of work. Third, although the payment is so little and there is doubt about data quality from this platform, there is little evidence that data from such an Internet-based environment is of lower quality than data from collected from a subject pool. Fourth,
researchers have discussions that MTurk can help enhance the internal validity because it avoids the interaction between experimenters and participants.

### 3.3.1 Experiment Design

The experiment is consisted of two conditions: a questionnaire and the *LinkIT* game. We use a climate change context in this study. A pre-determined “expert” mental model, provided by Bostrom [52], which consisted of 57 nodes guided the generation of questions in the questionnaire and puzzles in *LinkIT*. For the convenience of discussion, we use the word puzzle to denote a response prompt in both experimental conditions. After answering some demographic information related questions, each participant was randomly assigned to one of the two conditions and was required to continue the activity for at least 15 minutes.

![Figure 3.1. An example puzzle in the questionnaire to study the external validity of *LinkIT* as an elicitation tool for risk relations. Each puzzle presents a pair of risks and a list of options describing their relations. Participants are asked to choose one of these options that match with their own perception.](image)

Figure 3.1 shows an example puzzle in the questionnaire. This pairwise discussion/comparison format is widely used for similarity ratings [48] and relation elicitations [53]. The questionnaire puzzle first provides 2 risk events A and B, along with a
brief description of each event. Participants are asked to make a single selection from 5 choices: (1) A may affect B, (2) B may affect A, (3) A may affect B and B may affect A, (4) No such an affecting relationship exist between A and B, and (5) I dont know/Skip this question. With 57 nodes, we have a total of 1,596 pairs and 3,192 possible relations. For each participant, the puzzles are presented one by one with a random sequence. It is impossible for any participant to respond to all puzzles within 15 minutes. Thus, we adopt a naive random puzzle selection algorithm during the experiment so that each puzzle has an equal probability to be presented and responded to.

To match the questionnaire design, we limit the LinkIT game to present puzzles containing 2 nodes, as shown in Figure 2.1. In this way, participants could respond to each pair with the same five choices provided in the questionnaire. Similarly, puzzles are presented to players at random. For the purpose of anti-cheating, we pair participants with computer agents without telling them their partners’ identities; thus, the game was similar to the duo version but with the mechanics of the solo version.

After engaging in their task for a minimum of 15 minutes, participants are then asked to complete a post-experience survey containing 20 statements that query participants’ perceived (1) interface and usability, (2) background knowledge and subject matter interests, (3) fun and challenge, and (4) recommendation willingness and overall judgment. Table 3.1 shows these 20 statements and the corresponding category. Participants are asked to rate their degree of agreement on the statements, using a liked scale ranging from 1 to 5, with 1 representing strong agreement and 5 representing strong disagreement.
Table 3.1. The 20 statements in the post-experience survey, covering 4 categories. Statements with * are cast in negative terms.

<table>
<thead>
<tr>
<th>Interface and usability</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• The experiment interface was aesthetically appealing.</td>
<td></td>
</tr>
<tr>
<td>• The experiment interface was easy to use.</td>
<td></td>
</tr>
<tr>
<td>• I did not know what to do in the experiment. *</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Background knowledge and subject matter interests</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• I do not have much knowledge about climate change. *</td>
<td></td>
</tr>
<tr>
<td>• The subject matter was interesting.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fun and challenge</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• I concentrated during the experiment.</td>
<td></td>
</tr>
<tr>
<td>• The activity in this experiment was fun.</td>
<td></td>
</tr>
<tr>
<td>• The activity in this experiment was engaging.</td>
<td></td>
</tr>
<tr>
<td>• The activity in this experiment was easy.</td>
<td></td>
</tr>
<tr>
<td>• The activity in this experiment was frustrating. *</td>
<td></td>
</tr>
<tr>
<td>• The activity in this experiment was educational.</td>
<td></td>
</tr>
<tr>
<td>• The activity in this experiment was challenging.</td>
<td></td>
</tr>
<tr>
<td>• The activity in this experiment was exciting.</td>
<td></td>
</tr>
<tr>
<td>• I felt that the time dragged on during the activity. *</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Recommendation willingness and overall judgment</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• I would recommend this activity to others.</td>
<td></td>
</tr>
<tr>
<td>• If the activity in this experiment were made available to me online, I would voluntarily participate on my own time.</td>
<td></td>
</tr>
<tr>
<td>• I felt my time spent on this activity was worthwhile.</td>
<td></td>
</tr>
<tr>
<td>• Monetary payments are necessary for people to participate in this study. *</td>
<td></td>
</tr>
<tr>
<td>• I do not want to participate in this kind of study again. *</td>
<td></td>
</tr>
<tr>
<td>• My friends would enjoy participating in this study.</td>
<td></td>
</tr>
</tbody>
</table>
3.3.2 Participants

We recruited 240 participants from Amazon Turk workers and each participant was assigned to play the game or respond to the questionnaire randomly. At the end, each group ended up with 120 participants. Published literature suggests the dependence of lay people’s risk perception on demographic characteristics, individual attitudes, and cultural and institutional affiliations [42]. In this study, we assume that an aggregated group mental model exists for each group and that the variances brought by these demographic factors are equally distributed in the two groups.

3.4 Results

We analyze the experiment data and test the three hypotheses, construct validity, productivity, and enjoyment, respectively.

3.4.1 Construct Validity

To test H1, we construct group mental models for the two participant groups and conduct network-level comparisons.

3.4.1.1 Influence diagrams and probabilistic networks

We use influence diagrams to represent the group mental models. One way to formalize the problem for quantitative analysis is to employ probabilistic networks [54]. The concept has been demonstrated as a successful method to represent knowledge structure for reasoning under uncertainty in many fields such as cognitive science [55] and expert systems [56]. In the literature, discussions are mainly limited to acyclic networks where no loops are allowed in the networks. However, when considering risks, retrospective
causal or affective relationships exist. For example, chemical explosions may lead to fire and endanger buildings. Meanwhile, buildings storing flammable chemical materials on fire may cause hazardous chemical explosions. Thus, the algorithms for probabilistic network construction and analysis in the literature are not applicable here.

We adopt a straightforward approach in this study. All nodes are pre-determined from the reference expert model. Each potential link is placed or removed according to participants’ aggregated agreement on the corresponding affective relation. For each relation (A may affect B), there are \( N_0 \) responses, and \( N_+ \) of them present agreement on the relation. Let \( p = N_+/N_0 \) represent the percentage of agreement on the link. We further use the Clopper-Pearson method to estimate \( p \) and the 95\% confidence intervals \([p_l, p_h]\) for \( p \) [57]. The subscripts \( l \) and \( h \) refer to low and high, respectively. In this way, a binomial distribution is adopted to formalize the problem.

Because no commonly accepted metric to measure the similarity or difference between two probabilistic networks exists, we define 3 approaches to do the comparison. Moreover, we define a “random game-play group” to simulate the case when all LinkIT participants play the game randomly, acting like a random bot.

### 3.4.1.2 Estimation-based significantly different links

For each link, the responses from each group help us estimate the probability \( p \) associated with the link and construct a confidence interval \([p_l, p_h]\). Further, we use superscripts \( L \) (for LinkIT) and \( Q \) (for questionnaire) to differentiate the two participant groups. An estimation-based significantly different link is defined as:

\[
p^L \notin [p^Q_l, p^Q_h] \quad \text{or} \quad p^Q \notin [p^L_l, p^L_h]
\]
Because of the large number of puzzles in the database and the limited number of participants, some puzzles have not been presented to participants in this study or only have been presented only few times. Without loss of generality, we only include those puzzles that have been responded by at least 3 times by both groups. In total 1864 puzzles satisfy this requirement. Among them, we find that only 153 (8.21%) are significantly different between the two groups. However, if the LinkIT group plays the game at random, a simulation shows that 529 (28.38%) links would be significantly different.

3.4.1.3 Difference function

We define a difference function \( f(p_0) \) as follows. The parameter \( p_0 \) is a non-negative number between 0 and 1 that helping us convert the probabilistic networks into deterministic networks. Using \( L \) and \( Q \) to denote the 2 group mental model networks, for each link \( i \rightarrow j \) in \( L \), if the estimated probability \( p_{ij} \geq p_0 \), we set \( L_{ij} = 1 \). Else, \( L_{ij} = 0 \). Similarly, we can construct \( Q \). The difference function is defined the percentage of links that \( L_{ij} \neq Q_{ij} \), i.e.,

\[
    f(p_0) = \frac{\sum_{i,j \neq i} |L_{ij} - Q_{ij}|}{n(n-1)}
\]

in which \( n(n-1) \) is the total number of possible links. Apparently, when \( p_0 = 0 \) or 1, we have \( f(p_0) = 0 \), because the two networks would be exactly the same (fully connected or fully separated). When \( p_0 \) is around 0.5, \( f(p_0) \) achieves its maximum. The lower this maximum, the similar the two networks are to each other. Figure 3.2 shows the plot of \( f(p_0) \) with difference \( p_0 \). We also simulated the case when the LinkIT group plays the game randomly and compared the resulting network with the questionnaire network. The plot shows that the two experiment networks would differ from each other by 15.77% at maximum, while a random game-play network would differ from the questionnaire network by 50.38% at most, a lot worse than the actual LinkIT experiment.
network.

![Difference Function vs. Probability Threshold](image)

**Figure 3.2.** The difference function $f(p_0)$ vs. $p_0$. The blue dots and line are the result for the real experiment *LinkIT* group vs. the questionnaire group, while the red ones are the result for the simulated random game-play group vs. the questionnaire group.

### 3.4.1.4 Categorization

In this approach, we define 4 categories for the links according to the estimated probability $p$ associated with each link, as shown in Table 3.2. Applying this categorization method to the three group mental models, we obtain the results shown in Table 3.3. The numbers are the number of puzzles falling into each category in each group.

**Table 3.2.** The categorization method for the links, $p$ is the estimated probability associated with each link.

<table>
<thead>
<tr>
<th>Category</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong agreement</td>
<td>$p \geq 0.75$</td>
</tr>
<tr>
<td>Weak agreement</td>
<td>$0.5 \leq p &lt; 0.75$</td>
</tr>
<tr>
<td>Weak disagreement</td>
<td>$0.25 \leq p &lt; 0.5$</td>
</tr>
<tr>
<td>Strong disagreement</td>
<td>$p &lt; 0.25$</td>
</tr>
</tbody>
</table>
Table 3.3. The categorization results for the three group mental models. The numbers in the cells represent the number of links falling into each category.

<table>
<thead>
<tr>
<th>Category</th>
<th>LinkIT group</th>
<th>Questionnaire group</th>
<th>Random group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong agreement</td>
<td>113</td>
<td>280</td>
<td>334</td>
</tr>
<tr>
<td>Weak agreement</td>
<td>474</td>
<td>512</td>
<td>743</td>
</tr>
<tr>
<td>Weak disagreement</td>
<td>568</td>
<td>542</td>
<td>571</td>
</tr>
<tr>
<td>Strong disagreement</td>
<td>709</td>
<td>530</td>
<td>216</td>
</tr>
</tbody>
</table>

Among the 1864 puzzles, 1131 (60.68\%) of them fall into the same category in the two experiment groups. If we define a category-based significantly different puzzle as one that is categorized as Strong agreement in one group and as Strong disagreement in the other group, no link is significantly different. On the other hand, considering a simulated random game-play group, only 518 (27.79\%) of the 1864 puzzles would fall into the same category as the questionnaire group does and 150 (8.05\%) puzzles would be significantly different.

From Table 3.3, we also find that the LinkIT group provides fewer links than the questionnaire group. This is consistent with our preliminary findings and the existence of the cheating phenomenon [30]. In LinkIT, clicking “No Link” takes only one click while drawing links would take at least 2 clicks, while in the questionnaire, both options would take the same number of clicks. Consequently, LinkIT players have a tendency to submit more “No Link”’s than questionnaire participants. This fact realizes a need of an adjustment strategy on the network construction for the LinkIT group.

3.4.1.5 Summary

The three approaches to compare the two group mental models have two implications. First, the difference between the two probabilistic networks is considerably small:

1. Only 8.21\% puzzles are significantly different from each other in the sense that
the estimated probabilities differ.

(2) The maximum number of different links is only 15% of the total links under a deterministic approximation.

(3) 66.31% links are categorized the same in the two groups and no link is categorized significantly differently.

(4) Compared to a random game-play group, the group mental model from the LinkIT participants is a lot less different from the questionnaire group mental model. Table 3.4 summarizes the statistics.

**Table 3.4.** The results of comparing the LinkIT group mental model with the questionnaire group mental model. Column 2 is based on real game-play data and column 3 is based on a random game-play simulation.

<table>
<thead>
<tr>
<th></th>
<th>Real game-play group</th>
<th>Random game-play group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Significantly different links</td>
<td>153 (8.21%)</td>
<td>529 (28.38%)</td>
</tr>
<tr>
<td>Maximum value of the difference function</td>
<td>15.77%</td>
<td>50.38%</td>
</tr>
<tr>
<td>Links with significantly different categories</td>
<td>0 (0%)</td>
<td>150 (8.05%)</td>
</tr>
<tr>
<td>Links with the same category</td>
<td>1131 (60.68%)</td>
<td>518 (27.79%)</td>
</tr>
</tbody>
</table>

### 3.4.2 Productivity

To examine the productivity or the efficiency of the LinkIT game and to test H2, we compare the average number of puzzles each participant has completed in a unit time in the LinkIT group and in the questionnaire group. Because all participants have continued the activity for about 15 minutes, we can simply compare the total number of puzzles, \( N \), each participant has completed in that 15 minutes. Table 3.5 summarizes the mean
and standard deviation of $N$ for the two groups. A two sample t-test shows that in general the LinkIT group responds to significantly more puzzles than the questionnaire group does ($N(\text{LinkIT}) > N(\text{Questionnaire}), t[238] = 8.67, p = 0.000$). Therefore H2 is positively supported.

Table 3.5. Mean and standard deviation of the total number of puzzles N each participant has completed in the experiment (15 minutes).

<table>
<thead>
<tr>
<th></th>
<th>LinkIT</th>
<th>Questionnaire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>88.56</td>
<td>47.93</td>
</tr>
<tr>
<td>SD</td>
<td>41.61</td>
<td>30.03</td>
</tr>
</tbody>
</table>

3.4.2.1 Summary

We conclude that LinkIT outperforms the questionnaire approach with a better productivity and H2 is supported. The game allows participants to respond to a lot more puzzles than the questionnaire does, given the same amount of time.

3.4.3 Enjoyment

The post-activity survey (Table 3.1) data help us evaluate participants’ user experience in the activity. The 20 statements span 4 categories: (1) interface and usability, (2) background knowledge and subject matter interests, (3) fun and challenge, and (4) recommendation willingness and overall judgment. Because some statements are cast in negative terms, we reverse the responses to retrieve positive correlation coefficients.

We first calculate the value of Cronbach’s $\alpha$ for each category in order to measure the internal consistency of the survey questions. For the first category, the $\alpha$ value is 0.63 and we find that the questionnaire group rates all statements significantly higher than the LinkIT group does, implying that the LinkIT game has some usability concern.
The second category has \( \alpha \) value 0.62 and no significant differences are found between the 2 participant groups. For the other two categories, the \( \alpha \) values are 0.76 and 0.84, respectively. Because they have more questions and the focus of this survey is to examine whether the LinkIT game is more enjoyable than the traditional questionnaire, we further conduct exploratory factor analysis on them separately and apply two sample t-tests on each factor obtained.

For fun and challenge, 2 factors are produced with eigenvalues greater than 1 and can explain 58.25% of the total variances. Table 3.6 shows the factor loadings. The first factor accounts for 40.30% of the total variance and 35.12% of the common variance. The factor loadings indicate that this factor mainly expresses participants’ perceived enjoyment/fun and thus can be labeled as “Fun”. The second factor accounts for 17.95% of the total variance and 13.29% of the common variance. Upon close examined of the factor loadings, we label this second factor as “Easiness”.

**Table 3.6.** Factor loadings of the 2 factors retrieved from an exploratory factor analysis on the fun and challenge category. * indicates that the statement is cast in negative terms and the responses are reversed for data analysis. The questionnaire group has significant higher ratings on the italic statements and the LinkIT group has significant higher ratings on the bold statements.

<table>
<thead>
<tr>
<th>Questions</th>
<th>Factor 1 (Fun)</th>
<th>Factor 2 (Easiness)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I concentrated during the experiment.</td>
<td>0.393</td>
<td>0.059</td>
</tr>
<tr>
<td><strong>The activity in this experiment was fun.</strong></td>
<td>0.859</td>
<td>0.133</td>
</tr>
<tr>
<td>The activity in this experiment was engaging.</td>
<td>0.802</td>
<td>0.097</td>
</tr>
<tr>
<td>The activity in this experiment was educational.</td>
<td>0.635</td>
<td>-0.042</td>
</tr>
<tr>
<td><strong>The activity in this experiment was exciting.</strong></td>
<td>0.764</td>
<td>0.118</td>
</tr>
<tr>
<td>I felt that the time dragged on during the activity.*</td>
<td>0.552</td>
<td>0.003</td>
</tr>
<tr>
<td>The activity in this experiment was challenging.</td>
<td>0.543</td>
<td>-0.484</td>
</tr>
<tr>
<td><strong>The activity in this experiment was easy.</strong></td>
<td>-0.005</td>
<td>0.857</td>
</tr>
<tr>
<td><strong>The activity in this experiment was frustrating.</strong></td>
<td>0.203</td>
<td>0.425</td>
</tr>
</tbody>
</table>

For recommendation willingness and overall judgment, only 1 factor is produced
with eigenvalue greater than 1, explaining 56.63% of the total variance and 49.53%
of the common variance. Table 3.7 shows the factor loadings and we call the factor“Impression”.

**Table 3.7.** Factor loadings of the 1 factor produced from an exploratory factor analysis on the recommendation willingness and overall judgment category. * indicates that the statement is cast in negative terms and the responses are reversed for data analysis. The questionnaire group has significant higher ratings on the italic statements.

<table>
<thead>
<tr>
<th>Questions</th>
<th>Factor (Impression)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>I would recommend this activity to others.</em></td>
<td>0.868</td>
</tr>
<tr>
<td>If the activity in this experiment were made available to me online, I would voluntarily participate on my own time.</td>
<td>0.656</td>
</tr>
<tr>
<td>My friends would enjoy participating in this study.</td>
<td>0.807</td>
</tr>
<tr>
<td><em>I do not want to participate in this kind of study again.</em></td>
<td>0.620</td>
</tr>
<tr>
<td><em>I felt my time spent on this study was worthwhile.</em></td>
<td>0.798</td>
</tr>
<tr>
<td>Monetary payments are necessary for people to participate in this study.*</td>
<td>0.342</td>
</tr>
</tbody>
</table>

Because of the large sample size (120 in each group), a two sample t-test is appropriate to compare the two participant groups’ responses toward these three factors: Fun, Easiness, and Impression. Table 3.8 shows the mean and standard deviation of each factor in each group. The t-test results show that the questionnaire group considered the activity significantly easier ($t[238] = -4.09, p = 0.000$) and had higher overall judgment ratings ($t[238] = -2.18, p = 0.015$) than the *LinkIT* group did. With regards to Fun, the *LinkIT* group expresses non-significant higher ratings ($t[238] = 0.73, p = 0.766$).

We also conduct two sample t-tests on the responses to each statement separately and find that the *LinkIT* group has significant higher ratings toward 2 statements: (1) The activity in this experiment was fun, and (2) The activity in this experiment was exciting. On the other hand, the questionnaire group rates higher on 5 statements (italic in Table
Table 3.8. The comparison between the two participant groups, with regard to the 3 factors: fun, challenge, and impression. * indicates that there exist significant differences ($\alpha = 0.05$) between the two groups.

<table>
<thead>
<tr>
<th>Factor</th>
<th>LinkIT</th>
<th>Questionnaire</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Fun</td>
<td>16.68</td>
<td>3.33</td>
</tr>
<tr>
<td>Easiness*</td>
<td>2.99</td>
<td>1.41</td>
</tr>
<tr>
<td>Impression*</td>
<td>12.46</td>
<td>3.30</td>
</tr>
</tbody>
</table>

6 and 7). These results support the findings from the factor analysis.

3.4.3.1 Summary

In summary, the LinkIT game is perceived as more fun and more challenging (less easy), while participants have higher impression toward the questionnaire approach. Participants perceived the game to be of less usable (perhaps because questionnaires are inherently more familiar to participants than novel games) than the questionnaire and were not sure how to play the game. This may explain the low ratings on impression.

3.5 Discussion

Because of lack of a standard method to compare two probabilistic networks obtained via different techniques, our data and analysis do not directly say that LinkIT is externally valid. However, according to the three comparison approaches we adopted and the results shown in Table 3.4, we find that significant differences exist in only a small portion of links between the two participant groups. Moreover, when compared to what can be said about networks generated using a simulated random game-play group, the experiment LinkIT group generates a probabilistic network that are much more similar to the one generated by the questionnaire group.
A careful examination on the elicited networks suggests that the LinkIT group responds with “No Link” more frequently than the questionnaire group. Two reasons can be advanced to explain this. First, clicking “No Link” is much easier than drawing links between nodes. The former only needs one click at the center of the screen, while the latter needs at least two clicks at two different nodes (which are at different locations of the screen). This is similar to the framing effect: people’s decision making tends to depend on the way the options are presented [58]. Second, as found in preliminary studies, players tend to make consecutive clicks on “No Link” to cheat in game play [30]. Our anti-cheating algorithms can ease this problem but cannot prevent this cheating strategy completely.

The existence of cheating behaviors contributes to the differences between the two group mental models and affects our testing of H1. We propose several potential solutions for a more restrictive study. An intuitive thinking is to require two clicks for “No Link”’s instead of just one click. However, this might hinder the smoothness of game play experiences and creates usability problems. When constructing the group mental model from game players’ responses, we could put higher weights on links and lower ones on “No Link”’s to balance the framing effect. Also, we can build players’ credibility record that measures players’ honesty on playing the game. The network construction can be adjusted accordingly. At last, more advanced anti-cheating algorithms are possible to encourage players to provide honest responses in game play.

The results also show that the LinkIT game outperforms the traditional questionnaire in terms of both productivity and fun. Given the same amount of time, LinkIT can elicits more data by presenting more puzzles than the questionnaire and participants enjoy playing the game more than responding to the questionnaire. Also, the game is perceived as more challenging. However, the current game design may have some usability issues and some participants commented that they “did not know what to do
in the experiment”. This explains the lower ratings of the overall judgment of the game
than the questionnaire and suggests a need for user-centered interface design to improve
the game. These findings highlight the advantages of utilizing a gaming environment
for risk perception data elicitation.

This study has some limitations. First, we have only compared the LinkIT game
with a traditional questionnaire, assuming the existence of an expert model and a corre-
sponding set of risk events that can be used as a basis for generating puzzles. In reality,
we have not examined if such a set of events can be elicited from a limited number of
interviews on lay people or another gaming environment. The validity and comprehen-
siveness of these items are not guaranteed either. Thus, a more comprehensive study is
to apply the game in a systematic mental models approach and evaluate its performance,
or perhaps develop a series of interacting gaming system that, when combined, can fully
implement the entirety of the mental models approach.

Second, participants were paid and were required to participate for a fixed amount
of time. This differs from a usual risk perception case, in which participants volunteer
to respond to questionnaires at their own interests. This also detracts from the game
mindset; games are generally voluntary activities that people engage in for personal
enjoyment; to be paid to play seems to contradict what “play” means. An extension of
the current experiment is thus to distribute a web URL to the public and call for voluntary
participation without any stipulations on how to play (e.g., if cheating makes the game
fun, then let players cheat so long as the data is flagged as suspect). An interesting
study is to examine how monetary motivation and required participation time affects
peoples’ game play and their perceived gaming experiences. The data elicited from this
involuntary game play might differ from the data from a real “ludic elicitation” as well.

Third, the game itself relied on a naive puzzle selection algorithm for choosing which
puzzles to present to players and whom to pair with whom, whether human or bot. A
more sophisticated approach can enhance the gain, or the amount of information captured per puzzle played, by integrating intelligent puzzle selection strategies that takes into account the research questions that need answering, what puzzles have already been answered and by whom, who is playing in terms of demographics and credibility, and real-time dynamic player incentives that encourage continued game play. This problem set is huge and spans the interest of multiple research domains, and close attention to progress in this area will surely lead to more efficient ludic elicitation systems.

3.6 Conclusion

In this study, we examined the external validity, productivity, and enjoyment of the LinkIT game as a new ludic elicitation tool for eliciting lay peoples’ perceptions of the relationship between risks in different contexts and hence facilitating risk communication. A traditional questionnaire approach was adopted as a comparison to the game. Given the same set of risk events, experimental data suggested that the two approaches produced two group mental models represented as probabilistic directed networks that, when compared to one another, did not suffer from many significant differences. Our analysis did not provide enough evidence to fully support the external validity of the LinkIT game; however, it does enhance our confidence in the game’s value as a ludic elicitation tool. We further demonstrated that the productivity of LinkIT was significantly better than that of the questionnaire. Also, more participants playing the game reported that the activity was fun when compared to those responding to the questionnaire. However, the interface design of the current game was not perfect, and in fact, was a point of confusion in the game play. The game was also rated as more difficult than the questionnaire.

The results of this study have implications for future work. We have discussed the
application of LinkIT for eliciting lay peoples understanding of risk. Meanwhile, the
game can help elicit experts’ mental models, the first step of the mental models ap-
proach for risk communication. Generally, experts have more background knowledge
and higher interest on the risk context than lay people. As a result, their attitudes toward
the LinkIT game might be different and it is worthwhile to examine this application.
Also, according to LinkIT participants, noticing that they are actually playing with com-
puter agents was to some extent discouraging and affected their perception of enjoyment.
We are expecting to study the validity of the duo version of the game as an elicitation
tool and to investigate what effect social positioning has on the game’s productivity and
enjoyment. At last, in addition to supporting the communication between experts and
lay people in the long run, LinkIT also creates an opportunity for experts to educate lay
people by pairing lay people with experts to play the game together.

3.7 Another Application: Educational Performance

In another study, we found that LinkIT has the potential to serve as an educational game
[59]. By pairing human players with robot players who act like experts, human players
can learn experts’ knowledge on risk relations during the game play. By evaluating
players’ performances on a post-experience test, we concluded that the game performs
at least as well as a handout does. Again, the game is rated as more entertaining than
the traditional approach. Another interesting and possibly more important finding is that
the game has a significant improvement on the learning experience of those participants
who are initially interested in the learning context, while the knowledge handout does
not. This implies potential application of the game to current education systems for the
purpose of interest boost.
Chapter 4

Rank Elicitation: SortIT Elicits Ranks Data

The SortIT game is designed to elicit preferences by presenting puzzles in the form of multiple-choice questions. One intuitive application is to elicit ranks data. This chapter presents an experimental study that explores the construct validity and performance of SortIT as a rank elicitation tool.

4.1 Introduction

In decision analysis, ranks are widely used by expressing the preferences and/or evaluations of a group of raters/judges towards a set of items. For example, recommendation systems create ranks for products or services according to customer preferences and interests [60, 61]. In software development projects, prioritizing software requirements is a critical element of requirements engineering contributing towards making good decisions for software systems [62, 63]. In risk communication, experts are interested in how the general public perceive the relative riskiness (outrage) of different risks [64]. In
workplace environments, ranks are a very common and prominent feature as an evaluation approach [65]. In everyday life, people also make decisions on the basis of ranks, such as choosing a job offer, purchasing a house, and planning a vocation. Thus, rank elicitation is conducted in a variety of contexts. In many cases, either no sufficient data exist to create ranks, or the nature of the problem requires subjective inputs. Eliciting ranks from domain experts or the general public is thereby an important element of decision analysis [66].

While many publications study the analysis and aggregation of ranks, little attention is paid to methods for rank elicitation. A basic approach is to assign a numeric score/magnitude or ratio scale to each item in the candidate set, according to a specific criterion, and generate the rank ordering by the scores. An alternative is to first define several classes or categories of importance and then assign each item into the categories. A third one, the pairwise comparison approach is also frequently used for rank elicitation [60, 67, 68]. In each evaluation, one pair of items are compared against each other. The overall rank ordering is derived from the evaluation results. Some publications compare the results of different elicitation methods and find that they usually produce similar overall ranks [69, 70].

Surveys, questionnaires, and interviews are frequently used to elicit ranks or preferences data from domain experts and the general public [71, 72]. While interviews, preferentially structured, appear to be one of the most effective technique for elicitation purposes [9], they are usually too time and labor consuming to be applied over a large number of participants. The pairwise comparison approach is most frequently adopted in surveys and questionnaires. However, it presents a scaling issue. Let $N$ denote the number of items in the list. The number of pairs to be investigated is thus $N(N - 1)/2 \sim O(N^2)$. When $N$ is medium to large, it is not feasible to ask participants to compare all pairs. For example, 25 items will require 300 pairwise comparisons for a
comprehensive coverage. Some studies are thus devoted to the design of questionnaire queries with the purpose of minimizing the number of questions to be asked and therefore minimizing the data collection cost [73], such as the incremental vote elicitation method [74].

A second disadvantage of the traditional approaches for rank elicitation is the repeating nature of the task that kills participants’ patience and worsens the user experience. In this study, user experience is referred to as participants' enjoyment and fun in the process of conducting the data elicitation task [75]. Innovated by the concept of Games With A Purpose (GWAP), we propose to use a game approach to turn this controlled task to an open use situation [75], with the purpose of improving user experience.

Human computation is a field that studies the employment of human’s computing ability to solve complex or time-consuming problems that computers cannot easily solve in a reasonable time [14]. Games present a potential platform for human computation in an entertaining manner. Pioneered by Luis von Ahn and his colleagues, many games with a purpose (GWAP) systems have been developed in recent years. For example, the ESP game elicits high-quality image tags from the general public [20]; the CityExplorer game is a location-based mobile game and contributes to the collection of geospatial data [22]. Time-intensive or otherwise tedious tasks are crowd-sourced to the players and solved by human power. Thus, the “serious purpose” of a GWAP, that is, the designer’s intent for creating the game [23], is to incentivize productive work via the allure of fun. A number of publications have demonstrated the successes of these systems in fulfilling the serious purpose while providing an enjoyable experience for participants [20, 22].

We have developed the SortIT game to examine the application of GWAP for rank elicitation through the pairwise comparison approach. It forms the traditional ranking and comparison task as a play activity and thereby encourages continuous and enjoyable participation. This game differs from the existing Matchin game, which is designed to
elicit user preferences, specifically, image preferences, by pairwise comparisons as well. The SortIT game asks players to make choices for a certain question, while the Matchin game does not. This feature extends the flexibility of the SortIT game and makes it adaptable to most comparison-based data elicitation tasks, without limited to preference elicitation tasks. Further, because essentially the puzzles in SortIT is in the form of multiple-choice questions, the game is not limited to pairwise comparisons either. Another difference is that, with a very simple puzzle-game format, the SortIT game allows researchers to focus on the fundamental question: does the gaming platform provide a valid and effective environment for rank data elicitation tasks?

This question is important because the accuracy and truthfulness of the collected data from a gaming approach are not guaranteed and most times are plausible. The incentives provided in a gaming approach, such as winning the game, fighting with monsters, and collaborating with others, might distract players’ attention on the elicitation task and affect the data provided by game players. Thus, our focus of this study is to explore the validity of the gaming approach for rank elicitation.

This study explores the validity of this gaming approach for rank elicitation as well as its effectiveness in improving user experience. We conduct two experimental studies to examine two use situations respectively. In the first study, game players are the general public and we compare the game with traditional paper surveys in eliciting ranks from them. In the second study, game players are domain experts and we compare the ranks elicited from the game with those derived from real data. The findings in the two studies positively support the validity of this gaming approach and demonstrate the improvement in user experience.

The main contributions of this study are three-fold. First, we have developed a casual computer-based game that supports rank elicitation. It improves user experience by making the task more enjoyable and more durable. Also, it mitigates the scaling issue
brought by pairwise comparisons. Second, it generalizes the ranking task in one casual
game and supports knowledge elicitation from both domain experts and the general pub-
lic. Third, this is the first study on the validity of a gaming approach for rank elicitation.
Although many GWAP systems have been built and experimented, few studies focus on
the examination of construct validity. This study thus also contributes to the validity of
human computation games for data elicitation purposes in general.

The remainder of this chapter is organized as follows. Section 4.2 explains how
the SortIT game supports rank elicitation through a pairwise comparison approach and
reviews methods for evaluating ranking correlation, which will be used to evaluate the
validity of this GWAP application. Section 4.3 introduces the first experimental study
on this application, eliciting ranks from the general public. The experiment design are
described in detail, followed by the findings. Section 4.4 is a summary of our second
experiment study for rank elicitation with the participants to be domain experts.

4.2 SortIT for Rank Elicitation

As shown in Figure 2.6, the SortIT game provides an opportunity of leveraging players’
ranking abilities through the game play and a pairwise comparison approach. Because
the game proceeds in a faster manner than traditional questionnaires and surveys, we
expect this game to mediate the scaling issue brought by the pairwise comparison ap-
proach. Furthermore, because the game forms the rank elicitation task as a play activity,
we expect participants to enjoy the experience more than traditional survey experiences.

4.2.1 Create Ranks from Pairwise Comparisons

Several algorithms are available for creating ranks from pairwise comparison results. In
this study, we adopt the simple EWR approach and a relatively more complicated one,
the Relative SVD approach.

4.2.1.1 EWR

EWR stands for Empirical Winning Rate and this method assigns a simple measure, called winning rate, to each item in a list and builds the ranking order by sorting the winning rates. The EWR of a single item is the ratio between the out degree and the total degree, or the number of times it succeeds in a pairwise comparison and the total number of times it appears in a pairwise comparison.

\[ f_{EWR}(i) = \frac{\text{deg}^+(i)}{\text{deg}(i)} \]

Although this approach is easy to understand and implement, it has several limitations. First, if the total degree of an item is small, this metric is not a representative measure anymore. Second, in this approach, winning against a bad item is counted as the same as winning against a good item. It thus does not take into consideration the quality of pairwise comparison results. Further, it does not contain any user model and assign equal weights to all users. This might cause unfair results, especially considering the fact that SortIT is a gaming approach.

4.2.1.2 Relative SVD

The Relative SVD algorithm is proposed to calculate user preferences on images from pairwise comparison results [28]. It eliminates the limitations inherent in the EWR method by introducing (1) multiple features for each item \( V \) and each user \( U \) and (2) an iterative approach to minimize total sum of square errors. The error term is defined for each triplet \((i, j, k)\) where \( i \) is the user and \( j \) is the winning item against \( k \) in a pairwise
comparison. For each \((i, j, k)\) in the training data, the error is

\[ e_{ijk} = 1 - (\text{likes}(u_i, v_j) - \text{likes}(u_i, v_k)) \]

in which \(\text{likes}(u_i, v_j)\) defines the degree to which user \(i\) “likes” item \(j\) and

\[ \text{likes}(u_i, v_j) = u_i^T v_j \]

A more detailed description of the algorithm could be find in [28].

### 4.2.2 Ranking Correlation

Ranking correlation measures the degree of similarity between two different sets of ranks of the same set of items. Several methods are available to measure ranking correlation, such as the de facto metric Kendall’s \(\tau\), an intuitive measure of ranking correlation [76] and the average precision correlation coefficient [77].

#### 4.2.2.1 Kendall’s \(\tau\)

Denote the number of concordant pairs, which are in the same order in both rankings, as \(P\) and the number of discordant pairs, in reverse order, as \(Q\). Kendall’s \(\tau\) is defined as the ratio

\[ \tau = \frac{P - Q}{P + Q} \]

in which \(P + Q\) is also the total number of pairs. If \(Q = 0\), \(\tau = 1\), representing the case that the two lists are identical. If \(P = 0\), \(\tau = -1\), representing the case that the two lists are reversed. In general, \(\tau\) will be between -1 and 1, with a higher number referring to higher similarity.

Although simple and easy to apply, Kendall’s \(\tau\) has several limitations. All pairs
are treated equally and statistically independent with each other [78]. It penalizes errors (differences) that occur at any part of the list equally as well [77], while in general the top of a list is more important. Thus, it only provides a high-level approximation of rank similarity.

### 4.2.2.2 Average precision correlation coefficient

To overcome the limitations, Yilmaz and Aslam propose an AP method that penalizes errors made towards the top of a ranking more than the errors towards the bottom [77].

For two lists \( L_1 \) and \( L_2 \), the ranking quality of \( L_1 \) with respect to \( L_2 \) is

\[
p = \frac{1}{N - 1} \sum_{i=2}^{N} \frac{C(i)}{(i - 1)}
\]

in which \( C(i) \) is the number of items above rank \( i \) in \( L_2 \) and correctly ranked in \( L_1 \) with respect to the item at rank \( i \) in \( L_2 \). This metric falls between 0 and 1, where 1 means identical ranks and 0 means reverse order. A correlation coefficient between the ranks similar to Kendall’s \( \tau \) is thus defined as

\[
\tau_{AP} = 2p - 1
\]

which falls between -1 and 1.

### 4.2.3 Research Questions

This study is to examine the construct validity of the SortIT game as a rank elicitation tool as well as its performance in improving user experience. The construct validity is measured by the quality of the ranks elicited from the game, which can be represented by the ranking correlation between the game-elicited ranks and the ranks from
an established method. When the participants are the general public with no domain knowledge on the subject, we expect the game elicits similar ranks as the more traditional survey approach does. When the participants are domain experts, we expect the game elicited ranks to correspond with real data. The second research question in this study is whether the game improves user experience in terms of user enjoyment in the rank elicitation task, compared to more traditional approaches such as surveys.

4.3 Eliciting Ranks from the General Public

In this experiment, we examine the performance of the SortIT game in eliciting ranks from the general public and compare it with traditional surveys.

4.3.1 Experiment Design and Participants

We use a risk preference context in this experiment and elicit participants’ risk perception in terms of the relative riskiness of different hazardous activities and technologies. The 30 activities and technologies listed in Slovic’s risk perception paper [79] and 10 others are used. We adopt a within-subject design and each participant is asked to conduct two activities: (1) playing the SortIT game and (2) responding to a survey. All puzzles/questions in the game and in the survey are drawn from the 40 risks and in the form of pairwise comparisons, as shown in Figure 4.1. Because the puzzles are in the same format, the only difference is that the game approach adds a game environment to the survey approach. To avoid the priming effect of one activity over the other, we randomly assign the sequence of the two activities for each participant.

After each activity, the participant is asked to answer a short user experience survey. The first part of the survey is consisted of 11 statements (Table 4.1) that query participants’ perceived (1) interface and usability and (2) fun and challenge. These
Figure 4.1. Example puzzles used to elicit the general public’s risk preferences. The upper one is a survey puzzle and the lower one is a SortIT game puzzle.

statements have been tested in our previous studies and provide an effective measure for participants’ enjoyment in the experience. Participants are asked to rate their degree of agreement on each statement, using a likert scale from 1 to 7, with 1 representing strong agreement and 7 representing strong disagreement. The second part of this survey queries participants’ strategy in responding to the puzzles. Four possible strategies are proposed: (1) guessing, (2) using knowledge and experience, (3) making random choices, and (4) guessing what the partner would choose (this applies to the game only). Participants rate their use of each strategy according to 5 scales: (1) All the time (2) Most times, (3) Sometimes, (4) Rarely, and (5) Never.

We have recruited 22 undergraduate students from Penn State University to participant in the experiment. All of them have a major in either Security and Risk Analysis or Information Sciences and Technology, aging from 18 to 21. 14 participants are male and 8 are female. Each participant is randomly assigned to either play the game first or do the survey first. Each way ends up with 11 participants.
Interface and usability
- The experiment interface was aesthetically appealing.
- The experiment interface was easy to use.
- I did not know what to do in the experiment. *

Fun and challenge
- The activity was educational.
- The activity was exciting.
- The activity was easy.
- The activity was challenging.
- The activity was frustrating. *
- The activity was interesting.
- The activity was fun.
- The activity was engaging.

Table 4.1. The 11 statements in the post-activity user experience survey, querying participants’ perceived (1) interface and usability and (2) fun and challenge. Statements with “*” are cast in negative terms.
4.3.2 Data Quality and Construct Validity

4.3.2.1 Cheating Behavior

For more accurate analysis on the construct validity, we take a closer look at each participant’s experiment data, focusing on their game play and survey responding strategies. A cheating behavior is defined as one that satisfies one of both of the two conditions.

- The response data display an apparent pattern, such as “all left clicks”.
- The response speed is too fast to be true. A reasonable minimum time required for each puzzle is 5 seconds.

Interestingly, participants displayed more obvious cheating behavior in the survey activity than in the game activity. 8 participants responded to the survey by making the same or random clicks for most of the time in the survey, while only 2 of them admitted this cheating behavior. On the other hand, no apparent cheating behavior is detected in the game. 2 participants reported that they guessed the puzzle most of the time in the game play. The other participants reported that they mainly used their knowledge and experience. This finding that users cheat less in the game than in the survey provides positive supports for two hypotheses: (1) the game maintains (if not increases) the quality of elicited data, and (2) the game environment maintains users’ patience longer and decreases cheating behavior.

4.3.2.2 Ranking Correlation

Published literature suggests the dependence of laypeople’s risk perception on demographic characteristics, individual attitudes, and cultural and institutional affiliations [79]. In this study, we assume that aggregated group ranks exists for the participants. After eliminating the data from cheating behavior, we find that the game and the survey
Table 4.2. The two ranking correlation coefficients, Kendall’s $\tau$ and APCC, between the ranks elicited from the two approaches. We use two methods to generate the group ranks: Relative SVD and EWR.

<table>
<thead>
<tr>
<th></th>
<th>Kendall’s $\tau$</th>
<th>APCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative SVD: Game vs. Survey</td>
<td>0.608</td>
<td>0.413</td>
</tr>
<tr>
<td>EWR: Game vs. Survey</td>
<td>0.587</td>
<td>0.412</td>
</tr>
<tr>
<td>Game: Relative SVD vs. EWR</td>
<td>0.844</td>
<td>0.797</td>
</tr>
<tr>
<td>Survey: Relative SVD vs. EWR</td>
<td>0.854</td>
<td>0.818</td>
</tr>
</tbody>
</table>

elicit 1015 and 1221 responses, respectively. To measure the quality of ranks elicited from the game, we generate group ranks from the approaches separately and calculate the ranking correlation coefficients between them.

Table 4.2 summarizes the results. The two ranks construction methods, EWR and Relative SVD, leads to similar results (high ranking correlation coefficients) for both groups. Further, the two approaches, the survey and the game, elicits similar ranks, with Kendall’s $\tau$ to be around 0.6 and APCC to be around 0.4. The correlation coefficients between two random set of ranks will be around 0. Thus, we conclude that the two set of ranks elicited from the two approaches correspond with each other, at least significantly better than random.

4.3.3 User Experience

Table 4.3 displays the mean and standard deviation of all participants’ perception on the two activities, with 1 representing strong agreement and 7 for strong disagreement. The game is rated as less easier to use, indicating possible usability issues. However, participants rated the game play activity to be more challenging (less easy), more exciting, more interesting, more fun, more engaging, and less frustrating. These ratings positively support our hypothesis that the game improves user experience in the rank elicitation task. By forming the traditionally time and labor consuming task as a play activity, users’ patience and interest are enhanced and maintained longer. Also, the
Table 4.3. The mean and standard deviation of all participants’ perception on the two activities. Items marked with * present significant differences between the two activities.

The game environment adds a challenging feeling to users although the puzzles are actually the same as those in the survey. According to the flow theory, an appropriate challenging level is required for users to enter the flow state [80]. These findings are also consistent with our previous study in which a game is used to elicit structural knowledge [32].

### 4.4 Eliciting Ranks from Domain Experts

In this second experiment, the participants are domain experts and we compare the ranks elicited from the SortIT game against a real dataset.

#### 4.4.1 Experiment Design and Participants

We use the emergency medical dispatch (EMD) context in this experiment and ask participants to make pairwise comparisons on the relative severity of EMD categories in the SortIT game play. The categories are extracted from a real data set of EMD events in the Centre County area in Pennsylvania, with patients’ identifiable information removed.
A data analysis is conducted to identify common EMD categories that have at least 50 occurrences. A seriousness score/metric is defined as a function of patient conditions (minor, moderate, or life threatening). A ranked list of 27 common EMD categories is thus derived, with their relative seriousness. We refer to this list as data-driven ranks in the rest of this paper.

12 EMD service personnel in this region are invited to participate and play the game for about 20 minutes. Because participants are considered domain experts and the play time is relatively longer, we increase the game puzzle’s difficulty by including three choices in the puzzles after players have played the game for a while. Figure 4.2 shows such a puzzle. After the game play activity, we ask the same play strategy question as in the first experiment. Because this experiment does not have a control group nor another comparison approach, it does not include a user experience survey either.

**Figure 4.2.** The three-choice puzzle in the SortIT game under the EMD context.

### 4.4.2 Construct Validity

Because a real data set is available, it opens an opportunity to measure individual participant’s performance in responding to the puzzles in the game.
4.4.2.1 Players’ Performance

12 participants have played the game and responded to a total of 1658 puzzles, 1071 of which are supplied with 2 choices and the rest 587 puzzles are 3-choice questions. For each puzzle, there is one correct response according to the data-driven ranks. We calculate the correct response rate for each participant as well as the correct response rate by chance – when the participant plays the game by making random clicks. Figure 4.3 shows the real correct response rate and that by chance. We further conduct a one tail $t$-test for the paired data with the alternative hypothesis to be

$$H_a : \text{Players' correct response rate is better than chance.}$$

The returned $p$ value is 0.000. Thereby we conclude that players perform significantly better than chance.

Moreover, we examine participants’ responses to the play strategy question. All participants responded that they used their knowledge and experiences all or most of the time. Few participants guessed occasionally. Before the game play, we only primed participants that it would be a casual game play – without emphasizing that the data collected would be used to create ranks. Therefore, this finding supports the usefulness of the output-agreement game mechanism in eliciting high quality data.

4.4.2.2 Ranking Correlation

We use the Relative SVD algorithm and the EWR method to create two sets of ranks based on $SortIT$ game play data. The two sets are found to be highly correlated with each other and have high similarity scores. We then compare them with those derived from real data to examine the quality of the game-elicited ranks, using Kendall’s $\tau$ and the average precision correlation coefficient. Table 4.4 summarizes the comparison results.
**Figure 4.3.** Players’ correct response rate to *SortIT* puzzles vs. correct response rate by chance. Use box plot.
<table>
<thead>
<tr>
<th></th>
<th>Kendall’s $\tau$</th>
<th>APCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative SVD vs. EWR</td>
<td>0.960</td>
<td>0.925</td>
</tr>
<tr>
<td>Relative SVD vs. Data</td>
<td>0.755</td>
<td>0.494</td>
</tr>
<tr>
<td>EWR vs. Data</td>
<td>0.715</td>
<td>0.503</td>
</tr>
<tr>
<td>Random vs. Data $\sim N(\mu, \sigma)$</td>
<td>$N(0.327, 0.104)$</td>
<td>$N(0.006, 0.117)$</td>
</tr>
</tbody>
</table>

Table 4.4. Comparing the ranks generated from the game and those derived from real data using Kendall’s $\tau$ and the average precision correlation coefficient.

Further, because there is no standard or threshold on determining whether the game elicited ranks are satisfactory, we measure the similarity between a set of random ranks with the data-driven ranks, which follows a normal distribution. Statistical tests show that the game-elicited ranks are significantly better than random-generated ones in the sense that the former correlates with the data-driven ranks significantly more, i.e., with a significantly higher similarity score, than the latter ($p = 0.000$). This applies to both ranking correlation metrics.

### 4.5 Discussion

The two experiments demonstrate positive support for the construct validity of SortIT as a rank elicitation tool. In the first experiment, the ranks elicited from the game are in accordance with the ranks elicited from the survey. In the second experiment, players perform significantly better than chance and they respond to puzzles mainly according to their knowledge and experience. The aggregated game-elicited ranks also demonstrate high ranking correlation coefficients with the data-driven ranks. There are several facts that contribute to the small difference between the two set of ranks. First, there is no golden standard for the ranks of the risks in the first experiment or the EMD categories in the second experiment. Each person has his own perception even if the participant is a domain expert and thus the ranks elicited from any approach would contain subjective judgment. Individual differences contribute to minor conflicts in the ranks. Second,
both experiments do not have a large number of participants and each participant has only spent a small amount of time in the study. Thus the collected data might not be enough to construct the convergent group level ranks. Third, the game elicits ranks through pairwise comparisons. The algorithms we use to create ranks based on pairwise comparison results might not be the best fit for this application. Therefore, although there is no absolute evidence that the game elicits “high quality” ranks, the findings we obtained in this study still provide positive support on the validity of this gaming approach to elicit ranks.

A more important finding is that the game improves user experience in the rank elicitation task. Participants rated the game play activity to be more interesting, exciting, fun, engaging and less frustrating than the traditional survey activity. Also, the game adds an appropriate challenge feature to the activity and engages users more. Furthermore, the game decreases cheating behavior, compared with the traditional surveys. The fact that participants rated the game to be less easy to use brings up a potential usability issue of the current game design, meanwhile opens a new opportunity for even more improved user experience.

These findings are encouraging news to the application of games with a purpose in rank elicitation and other data elicitation tasks. In this rank elicitation task, the gaming approach eliminates the scaling issue of pairwise comparisons and improves participants’ user experience, while maintaining the quality of elicited data. It also brings some other advantages. First, the accessibility and convenience of a browser-based game supports the contributions from a large number of participants, including both domain experts and the laypeople. Second, it supports the “communication” and “collaboration” of group work (by the debriefing stage and the output-agreement game mechanism). Meanwhile, it eliminates undesired group dynamics in traditional workshop settings, such as an over-abundance of unfocused and rambling discussion and by counterpro-
ductive group dynamics [81]. Third, it enables easy separation of ideas from different
groups and statistical comparisons between competing ideas, because each single piece
of game play data (responses to puzzles) are recorded. Moreover, the most important
advantage might be the flexibility and adaptivity of the system. The puzzles could be
selected from any predetermined data set. Thus, theoretically speaking, the game is ap-
licable to any rank elicitation task. Changing the question prompt would help evaluate
user preferences under different criteria.

It is important to notice that SortIT only serves as a candidate application for this
gaming approach. Both the background story and puzzle formats are simple enough for
rapid production and experimentation. After having built the validity of the gaming ap-
proach, a more sophisticated game should be developed to ensure (1) fun, (2) fluent user
experience, and (3) successful data elicitation. In addition to the game design, future
work could also focus on improving the efficiency of this gaming approach. For exam-
ple, an intelligent and adaptive puzzle selection algorithm presents “smart” questions to
game players and thus increases the efficiency of the rank elicitation task.

4.6 Conclusion

Ranks play an important role in many systems and dimensions. In some fields, there is
no enough data to help derive ranks and therefore eliciting ranks from human being is
necessary. However, traditional approaches for this task, such as surveys and question-
naires, are usually time and labor consuming, not to mention the tiring user experience.
Pairwise comparison is commonly used to elicit ranks as well. However, it brings a
scaling issue because the number of pairs is proportional to the square of the number of
items to be ranked.

Innovated by the concept of games with a purpose, we propose a gaming approach to
elicit ranks through pairwise comparisons. In this paper we introduce a game construct that we have developed and introduce two experimental studies on this game. In the first study, the game is used to elicit ranks for 40 common risks from the general public and we compare its performance with a more traditional survey. In a between-group design, the game play activity is rated significantly more enjoyable than the survey activity. Also, the ranks elicited from the two approaches demonstrate high ranking correlation coefficients, indicating positive support for the construct validity of the game. In the second study, we elicit the ranks for 27 common EMD categories from 12 EMD related personnel who are considered as domain experts. The following findings are obtained. First, at the puzzle level, game players perform significantly better than chance and respond to puzzles with a significantly higher correct rate. Participants also reflect that they play the game based on their knowledge and experience for most/all of the time. Second, at the ranks level, the game-elicited ranks demonstrate a high similarity with the ranks derived from a real data set. Statistical tests also show that they correlate with the data-driven ranks significantly better than chance. Therefore, we conclude that the gaming approach is a valid tool for rank elicitation with significantly improved user experience.
Chapter 5

Probability Elicitation: SortIT Elicits Probabilities

In this chapter, we propose another application of the SortIT game supported by its flexibility in question types and the puzzle format. In this experimental study, objective probabilities are elicited from “domain experts”. The game is expected to be applicable for eliciting subjective probabilities as well.

5.1 Introduction

Decision support systems such as Bayesian networks usually need to model uncertainty in the form of probabilities. When no sufficient data exists to derive these probabilities, they are usually elicited from domain experts [82, 83]. Extensive psychological research studies have demonstrated that human beings, even experts, tend to encounter difficulties when assessing probabilities and provide poorly calibrated and biased assessments [84]. The goal of elicitation is to make it as easy as possible for domain experts to transfer their knowledge in probabilistic terms without requiring them to have
a deep understanding in probability theory [85]. A number of methods are designed to overcome or suppress this problem, such as probability wheels, gamble-like methods [86], and the use of numerical intervals and verbal anchors [82]. However, most of these methods are very time-consuming and presents a scaling issue and thus are infeasible for real world applications when thousands of probabilities are to be elicited [82]. Another main drawback of these methods is the repeating nature of the task – participants easily get bored and tired of the probability elicitation activity while these methods usually require a long time of participation.

We propose to apply the gaming approach for probability elicitation tasks, i.e., to form the elicitation activity as a play activity. The SortIT game serves as a potential platform for this application because of its flexibility in the format of the question prompt. One concern is that the probability elicitation task is more difficult than the previous two applications we have proposed (relation elicitation and rank elicitation). The game background might distract participants’ attention rather than help them focus. Thus, an experimental study is necessary to examine the validity of this application.

In this study, we only use a simple questionnaire puzzle format that use probability intervals for probability elicitation. After having established the applicability of the game approach, further studies could combine the game environment with more advanced elicitation formats such as probability wheels. A within-subject design is adopted in the experimental study to compare the performance of the game approach with the more traditional questionnaire approach. Again, we also ask participants to rate their perceived user experience for each activity.

The main contributions of this study are three-fold. First, we have developed a casual computer-based game that supports probability elicitation. It improves user experience by making the task more engaging, exciting, and fun. Second, it generalizes the probability elicitation task in one casual game and the game supports the application of
more advance probability elicitation tools such as probability wheels and other gambling methods. Third, this is the first experimental study on the validity of a gaming approach for probability elicitation. Although many GWAP systems have been built and experimented, few studies focus on the examination of construct validity. This study thus also contributes to the validity of human computation games for data elicitation purposes (that involve more difficult tasks) in general.

The rest of this chapter is organized as follows. Section 5.2 introduces the concept of a game approach for probability elicitation and explains how the SortIT game supports this application, followed by the research questions. Section 5.3 describes the experiment design and participants in this study. The findings are summarized and discussed in Section 5.4 and 5.5. The last section concludes the study.

5.2 A Game Approach for Probability Elicitation

The success of SortIT as a pairwise ranking tool implies that the game has the ability to elicit people’s perceptions in a multiple-choice question format. In addition to the flexibility in the question prompt, the number of choices can be adaptive as well. This flexibility can be utilized for probability elicitation from experts, using different probability information formats. One application is to elicit probability intervals. The granularity of the intervals corresponds with the difficulty of game puzzles. Figure 5.1 (top) shows an example scheme for this application. The first level of the game presents 3 choices for each probability question: [0, 0.4), [0.4, 0.6], and (0.6, 1]. After having narrowed down the range for the probability, e.g., [0, 0.4), we can proceed to the second level and use three smaller intervals: [0, 0.05), [0.05, 0.2), [0.2, 0.4). The first interval helps identify “extreme” events that has a small chance of happening. Similarly, an interval of (0.95, 1] helps identify “almost certain” events.
Another potential application is to use verbal probability expressions that correspond to numerical values, such as those shown in Figure 5.1 (bottom). However, having seven options in the game will make the game too challenging and even frustrating. One solution is to propose 3 choices in each puzzle and present new game puzzles adaptively according to previous game results. For example, in order to elicit the probability of an event, we could first ask players to choose from three options: (1) less than 50/50 (2) about 50/50, and (3) more than 50/50. After having narrowed down the range, e.g., less than 50/50, a new puzzle then asks players to choose from (1) expected, (2) probable, and (3) certain to describe the probability. Figure 5.1 (bottom) shows such an adaptive puzzle-selection strategy. Experimental tests need to be conducted to examine its applicability and reasonability.

5.2.1 Research Questions

This study is to examine the construct validity of the SortIT game as a probability elicitation tool as well as its performance in improving user experience. The construct validity is measured by the quality of the probabilities elicited from the game, which can be represented by the accuracy or correctness of the elicited probabilities, compared to those elicited from a more traditional approach. In order to measure user experience, we again use the experience survey that we have used in the rank elicitation study.

5.3 Methods

5.3.1 Experiment Design

We conduct an experimental study to explore the performance of SortIT as a probability elicitation tool and adopt a within group design. Each participant is asked to answer
Figure 5.1. An adaptive puzzle-selection strategy to elicit probability intervals (top) or verbal probability expressions (bottom) in SortIT, maintaining that each puzzle only contains three choices.
probability questions in two environments, the SortIT game and a questionnaire format, with a random sequence. All questions are in the form of multiple-choice questions and probability intervals are provided as the choices. In this way, participants do not need to calculate the exact probabilities. Figure 5.2 shows example puzzles in the two environments. Each activity is required for at least 12 minutes.

All probability puzzles are designed to be objective ones with an unambiguous answer, collected from or constructed according to college level probability textbooks. In total we have 90 puzzles with different difficulty levels. In order to minimize priming/memorization effect, the puzzles are randomly divided into two groups according to the difficulty levels and thereby two sets of puzzles are used for the two environments respectively.

After each activity, participants are asked to answer a short user experience survey. Table 5.1 shows the 11 statements in the survey, which measures participants’ (1) perception on the activity interface and usability and (2) perceived fun and challenge in the activity. A likert scale is used for each statement with 1 representing “I strongly disagree with the statement” and 7 representing “I strongly agree with the statement”.

### 5.3.2 Participants

22 Chinese students participated in this experiment, all of whom were pursuing a graduate degree at Penn State, majoring in one of the following majors: Computer Science, Information Science and Technology, Mathematics, Physics, or Biology. Thus, an assumption that all participants are “experts” in probability questions could be made. 11 participants played the game before responding to the probability questionnaire while the other 11 conducted the two activities in the reverse order.
Figure 5.2. Example probability questions in the SortIT game (above) and in the questionnaire (below). Participants are asked to answer these multiple-choice questions in the probability elicitation study.
Table 5.1. The 11 statements in the post-experience survey, covering interface and perceived fun. Statements with * are cast in negative terms.

<table>
<thead>
<tr>
<th>Interface and usability</th>
</tr>
</thead>
<tbody>
<tr>
<td>• The experiment interface was aesthetically appealing.</td>
</tr>
<tr>
<td>• The experiment interface was easy to use.</td>
</tr>
<tr>
<td>• I did not know what to do in the experiment. *</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fun and challenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>• I concentrated during the experiment.</td>
</tr>
<tr>
<td>• The activity in this experiment was fun.</td>
</tr>
<tr>
<td>• The activity in this experiment was engaging.</td>
</tr>
<tr>
<td>• The activity in this experiment was easy.</td>
</tr>
<tr>
<td>• The activity in this experiment was frustrating. *</td>
</tr>
<tr>
<td>• The activity in this experiment was educational.</td>
</tr>
<tr>
<td>• The activity in this experiment was challenging.</td>
</tr>
<tr>
<td>• The activity in this experiment was exciting.</td>
</tr>
<tr>
<td>• I felt that the time dragged on during the activity. *</td>
</tr>
</tbody>
</table>
5.4 Results

We are interested in participants’ productivity, performance, and experience in the two probability elicitation activities: playing the game and responding to the questionnaire.

5.4.1 Productivity

On average, participants have spent similar amounts of time and answered similar numbers of puzzles in the two activities. Thus, no productivity difference is identified. The total numbers of puzzles across all 22 participants in the game and the questionnaire environments are 742 and 640, respectively. These two numbers provide us with good confidence on assessing participants’ performance.

5.4.2 User Performance

We examine participants’ correct rate on answering the probability questions in the two environments. Figure 5.3 shows the results and Table 5.2 summarizes the statistics. As shown, users’ average performance in the questionnaire environment is only a little bit better than that in the game environment (0.80 > 0.75). Further, a paired student t-test on the data returns $p = 0.06, t = 2.02, df = 20$. Because the $p$ value is more than 0.05, with 95% confidence, we accept the null hypothesis that participants reached the same performance in the two environments.

Table 5.2. Mean and standard deviation of participants’ correct rate of answering the probability puzzles in the two activities: SortIT game vs. questionnaire.

<table>
<thead>
<tr>
<th></th>
<th>SortIT</th>
<th>Questionnaire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct rate</td>
<td>0.75</td>
<td>0.80</td>
</tr>
<tr>
<td>SD</td>
<td>0.12</td>
<td>0.15</td>
</tr>
</tbody>
</table>
Another interesting exploration is the correct rate of each puzzle (although few of them are answered only a few times) in the two activities, as shown in Figure 5.4. The horizontal axis represents the puzzle indexes, which is also an indication of the puzzle difficulty level. Although the average correct rate of the questionnaire is a little bit higher than that of the game, a t-test again reveals that no significant difference exists between the two activities ($t = 1.94, df = 86, p = 0.06$). Further, we (roughly) categorize all puzzles into 5 levels, according to the time spent of two “experts” in answering them. Figure 5.5 shows the average correct rate of each level in the two activities.

Figure 5.3. Participants’ correct rate of answering the probability questions in the two activities: SortIT game vs. questionnaire.

Figure 5.4. The average correct rate of each puzzle in the two activities: SortIT game vs. questionnaire.
5.4.3 User Experience

Participants are asked to rate their degree of agreement on 11 statements (Table 5.1) in the two activities. Statistical tests result in significant differences in terms of (1) excitement, (2) fun, (3) engagement, and (4) interest and no significant differences in the other aspects. Table 5.3 summarizes the mean and standard deviation of these four dimensions in the two activities. Consequently, we conclude that the game improves participants’ user experience in the probability elicitation activity.

Table 5.3. Mean and standard deviation of participants’ rated excitement, fun, engagement, and interest in the two activities: SortIT game vs. questionnaire.

<table>
<thead>
<tr>
<th></th>
<th>SortIT</th>
<th>Questionnaire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>Excitement</td>
<td>4.24</td>
<td>1.71</td>
</tr>
<tr>
<td>Fun</td>
<td>4.62</td>
<td>1.69</td>
</tr>
<tr>
<td>Engagement</td>
<td>4.33</td>
<td>1.60</td>
</tr>
<tr>
<td>Interest</td>
<td>4.43</td>
<td>1.66</td>
</tr>
</tbody>
</table>
5.5 Discussion

According to our experimental data, the SortIT game maintains the productivity and user performance of traditional probability elicitation questionnaires. For objective probability puzzles, participants have completed about the same number of puzzles (slightly more) in the game play activity, compared to the questionnaire activity, within the same amount of time. They have also demonstrated similar correct rates in the two activities. This finding positively supports the external validity of the game approach for (objective) probability elicitation. Without losing generality, we expect the game to be applicable for subjective probability elicitation, based on this finding as well the findings in the rank elicitation study (Chapter 4) and the structure elicitation study (Chapter 3).

One challenge of probability elicitation tasks is that when group decision is required, it is usually difficult to achieve group level agreement. The SortIT game presents a potential to facilitate group communication on probability elicitation tasks because of the two-player game feature. Each player is allowed to present his/her own opinion on each puzzle. However, the winning of the game is based on their mutual understanding. The debriefing stage (Section 2.2.3) presents both players’ opinions and facilitates the communication.

The most important improvement brought by the game approach lies in the user experience, as we have found in the previous two studies. Our participants rated the game play activity to be significantly more enjoyable (in terms of engagement, interest, fun, and excitement) than the questionnaire activity, although the nature of both activities is probability elicitation. Because participants easily get bored in traditional elicitation tasks, this improved user experience has the potential to make the tasks more durable as well.
One limitation of this study is the participant size. More participants and extended participation would allow us to better examine the external validity and performance of the game approach. However, because basically all our three studies have the same research questions and the total number of participants is a large number, we are confident about the findings here. Another limitation is that all participants are graduate students at Penn State who originally come from China. Thus this study is limited to one particular group and future studies with other groups will help generalize our findings.

Consequently, we propose two future studies. First is the facilitation of the game approach on group level probability elicitation. This may bring some minor changes on the game mechanism. For example, more time should be devoted to more interactive group communication (the debriefing stage). Animations provide a useful tool for enhanced interaction. Another possible change is on the puzzle format and the winning condition. A continuous probability scale instead of several discrete choices can be used for participants to indicate their probability assessment. At the initial stage of the game, less challenge is usually desired, which can be realized by loosened winning condition. For example, if the two players’ selected probabilities are in the same range (or has a relatively small difference), the game proceeds with a match. As the game continues on, more challenges is introduced by shrunken probability ranges (or differences), used for winning judgements.

A second future study is to apply the game approach for probability elicitation in a more general sense. For example, both subjective and objective probability puzzles should be used and participants from different age, profession, nationality groups should be invited for longer time of participation. The current study employs computer agents to play as participants’ partners. When the number of participants is large enough to support the two-player game mode, we will have no need of building computer agents for the probability elicitation task. Another benefit of this is the minimization of designer-
participant interaction because computer agents are mostly based on designers’ understanding of the probability puzzles.

5.6 Conclusion

In this study, we examined a second application of the casual SortIT game as a probability elicitation tool. We conducted an experimental study to examine the external validity, productivity, and enjoyment of the game. A traditional questionnaire approach was adopted as a comparison to the game. Using two sets of probability puzzles with similar difficulty levels and formats, experimental data suggested that participants performed with similar correct rates of responding to the puzzles in the two environments. This finding positively supported the external validity of the game approach for probability elicitation. Further, participants rated the game play experience to be significantly more enjoyable than the questionnaire activity. We thus conclude that the game improves user experience of probability elicitation tasks without sacrificing performance.
Chapter 6

Future Application: Bayesian Network Elicitation

Combining the three studies introduced in Chapter 3, 4, and 5, we propose a game approach for Bayesian network elicitation. First, we can use the LinkIT game to elicit the relations between variables and then construct the Bayesian network structure. Second, the SortIT game can be used to elicit the conditional probabilities for the Bayesian network. Because of limited time and budget, we have not conducted experimental studies on this application yet. This chapter only proposes the application schema.

6.1 Bayesian Networks

A Bayesian network is a directed acyclic graph where the nodes represent propositional variables of interest and the directed edges (links) define direct informational or causal dependencies between the variables [87, 88]. Figure 6.1 shows a simple BN, containing five variables: Age, Occupation, Climate, Disease, and Symptom. The links represent causal dependencies. For example, the link from Disease to Symptom indicate that
Disease has a causal impact on Symptom. As shown, such a model provides a representation of expert knowledge and reasoning. In literature, some other names are used for BNs, such as probabilistic network [89], belief network [90], knowledge map [91], and so forth.

![Bayesian Network Example](image)

**Figure 6.1.** A simple Bayesian network example that has 5 variables and 4 relations, adapted from [92].

In such a diagram, the network structure is usually referred to as the qualitative part [93, 94]. One important aspect of BNs is the conditional independence assumption. One node is conditionally independent of any other node, given the value of all its parents. Thus, the probability distribution of any variable only depends on its parents. The quantitative part of BNs capture this dependence. It is the conditional probability distributions of each variable given the values of all its parents in the network, usually denoted as conditional probability tables, or CPT [93, 94].

BNs are widely used to reason under uncertainty, create expert systems, and supplement with other decision support tools [95, 89]. Its application spectrum has been spread widely in recently years, covering environmental modeling and management [89], medical diagnoses [96], text analysis [97], gene studies [98, 99], and so forth. For example, in computational biology, Friedman etc. use BNs to analyze expression data and to uncover gene/protein interactions and key biological features of cellular systems from
DNA hybridization arrays [98]. Some other tools and techniques of automated reasoning include the classical logics and calculi, fuzzy logic, and some less formal ad-hoc techniques.

6.2 Constructing Bayesian Networks

Constructing a BN for a domain of application involves three steps [100]. First, we need to identify the variables that are of interests, as well as their possible values. These variables constitute the node set for the network structure. The second step is to identify the dependence relations between the variables and to express these relations in a graphical representation, i.e., in a network structure. These two steps elicit the qualitative structure of a BN. The third step is to obtain all the conditional probability tables for the network. In the following, we refer to the first two steps as qualitative structural development, and the third one as the quantitative probability estimation. In practice, after building a BN, experts usually need to evaluate the structure and the probabilities. This research focuses on the construction process and at this current stage does not consider the evaluation step.

6.2.1 A Causal Mapping Approach

Nadkarni and Shenoy proposed a causal mapping approach to constructing BN structure, which is to first elicit causal maps from domain experts and then modify causal maps to construct BNs [53]. Causal maps, also called cause maps or cognitive maps, are directed graphs that capture the cause-effect relations between concepts in the context of a particular issue/event [53, 101], “characterized by a hierarchical structure which is most often in the form of a means/end graph” [102]. Thus, causal maps and Bayesian networks are both causal models. Similar to BNs, causal maps are frequently used to
make decisions [103].

**Constructing Causal Maps**

Hodgkinson, Maule, and Bown summarizes different ways in literature to constructing causal maps and focuses the comparison on two alternative approaches: pairwise evaluation of causal relationships and a freehand approach [101]. The pairwise evaluation approach is to ask participants to evaluate the relation between each pair of the variables. Although this method is undoubtedly a thorough procedure, it is also relatively time-consuming and labor-intensive, because the number of pairs is $O(N)$, in which $N$ is the number of variables in the context. In the second approach, participants are asked to represent their beliefs visually in a free manner and researchers use this self-reported data to construct causal maps. This approach is quicker to administer and has less requirement on the number of pairwise causal judgments. Conducted an experimental study on these two approaches, Hodgkinson etc. concluded that the more time-consuming and labor-intensive approach yielded relatively more elaborate maps while participants found the task less engaging, more difficult, and less representative [101]. The method proposed in this research is a pairwise evaluation approach. However, it is expected to ease the time and labor requirements on participants. Section ?? and ?? presents more details about this argument.

**From Causal Maps to Bayesian Networks**

Causal maps and BNs differ from each other in several ways [104, 53]. Here I only review those related in this research. The first difference lies at the conditional dependencies in the two structures. A network model can be either a dependence map (D-map) or an independence map (I-map). In a D-map, variables found to be connected are indeed dependent; while in an I-map, variables found to be separated are indeed conditionally
independent, given other variables [104]. A causal map is a D-map and the absence of a link between two variables does not imply a lack of dependence. However, in BNs, an absence of a link between two variables implies conditional independencies between them. Thus, BNs are I-maps. An example provided in the literature is shown in Figure 6.2, in which three links are added to make a causal map a D-map and an I-map.

![Causal map example](image)

**Figure 6.2.** An example showing that three links are added to make a causal map a D-map and an I-map. The new links change the conditional independence assumptions about variables in the map. This example is adapted from [53].

Second, causal maps do not distinguish “direct” and “indirect” links while the conditional independencies in BNs require such a distinction. This difference is shown in Figure 6.3. Third, causal maps allow the existence of circular loops and bi-directional relations, violating the acyclic nature of BNs. Two reasons may lead to circular loops in a causal map. One is coding mistakes and the other one is that they may represent dynamic relations between variables across multiple time frames.

In summary, to convert causal maps to BNs, we need to add missing links, remove indirect links, eliminate circular relations and bi-directional relations. In [53], the authors propose to use structured methods such as interviews and adjacency matrices to eliminate these biases. Figure 6.4 provides an illustration of the structured interviews for this purpose, in which experts are asked to specify the direct relation between pairs of variables. The adjacency matrix method is to ask experts to denote the relations between
Figure 6.3. An example showing that causal maps might contain indirect relations and they need to be eliminated to convert a causal map to a BN. This example is adapted from [53].

variables in a matrix in which rows are causes and columns are effects.

Figure 6.4. An illustration of the structured interview that elicits experts’ opinions on the causal relations between variables [53].

6.2.2 Quantitative Probabilities Elicitation

Bayesian probabilities differ from classical probabilities in the sense that the former are properties of a person who assigns the probabilities while the latter are physical
properties of the world. In other words, the Bayesian probability of an event is a person’s degree of belief in that event [105].

Because even a simple BN might have a very complex CPT structure and domain experts are usually reluctant to give exact numbers for probabilities, the elicitation of probabilities is recognized as a major obstacle for BN construction [93, 106]. It is also one of the most laborious tasks in building other decision-theoretic models. However, the elicitation of probabilities has so far received only moderate attention in decision-theoretic systems.

When the odds are not objectively measurable, a lot of people feel more comfortable with verbal probability expressions than with numbers [94]. Also, not all probabilities can be cast as exact numbers [107]. Many researchers thus study the use of other forms of probabilistic information, in addition to numerical and interval probabilities, such as estimates of order of magnitude, signs of influences and/or synergies, purely qualitative statements regarding the dependencies between variables. Although numbers advances words in the sense that numbers are precise, allow calculations, and have a fixed rank order, expressing probabilities verbally is very often perceived more natural, easier to understand, and better to convey the vagueness of subjective opinions [108].

For example, Druzdzel and van der Gaag [93] proposed a non-invasive approach to eliciting probabilities from domain experts by eliciting what ever probabilistic information experts are willing to state, whether qualitative or quantitative in nature. These information are then used to derive second-order probability distributions using a canonical form. In this approach, probability elicitation can be looked upon as constraining the distribution space (of all possible joint probability distributions) as much as possible.

Bonissone and Decker [109] examined the use of linguistic terms to help establish a certain granularity to facilitate knowledge elicitation. In this way, precision is traded off by decreased complexity. Renooig and Witteman [94] proposed to combine a verbal
expression scale with a numerical scale to help ease both the elicitation and communication of probabilities. Experimental studies were conducted to explore the matching between verbal expressions and numerical probabilities. Table 6.1 shows the final scale with seven categories of probability expressions with the corresponding numerical values. [106] also examined this scale correspondence.

Table 6.1. The matching between verbal expressions and numerical probabilities [94]

<table>
<thead>
<tr>
<th>Verbal Expression</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>certain</td>
<td>100%</td>
</tr>
<tr>
<td>probable</td>
<td>85%</td>
</tr>
<tr>
<td>expected</td>
<td>75%</td>
</tr>
<tr>
<td>fifty-fifty</td>
<td>50%</td>
</tr>
<tr>
<td>uncertain</td>
<td>25%</td>
</tr>
<tr>
<td>improbable</td>
<td>15%</td>
</tr>
<tr>
<td>impossible</td>
<td>0%</td>
</tr>
</tbody>
</table>

The use of imprecise probability measures has been widely accepted and adopted by many researchers. In particular, the concept of Quasi-Bayesian networks or credal networks is to associate convex sets of probability measures, which are imprecise, with direct acyclic networks [107]. Theories and algorithms have been developed for inferences in such networks. Moreover, the inference procedures in a BN are more sensitive to the qualitative structure than the quantitative probabilities associated with the structure [110]. This finding supports the use of imprecise probabilities.

Many researchers have also studied the sensitivity of BNs of imprecise probabilities and found that even highly imprecise input probabilities may not have a significant impact on the performance of BNs [111]. These findings indicate the BNs are a practical representation without requiring undue precision.
6.2.3 User Experiences

Although many researchers and publications are devoted to the elicitation of BNs from domain experts, nearly no one has studied the experts’ experiences in this process. Usually, groups of experts need to get together or communicate through online media to have iterations of discussions to determine the nodes, the relations, and the probabilities. The process is usually long, tedious, and difficult – experts need to keep a record of what have concluded so far and reach agreements to proceed to next steps. Furthermore, experts need to take charge of data encoding that adds more difficulties to the task.

Pfautz, Cox, and Catto, etc. proposed several user-centered methods for the construction of BNs, focusing on how experts tend to express their reasoning about the domain [112]. However, they are only considering potential improvements on dynamic user interfaces that are interactively created with or by domain experts. Another study has also proposed to use interactive graphical interfaces to facilitate the elicitation of probabilities [113].

We propose to use an innovative platform, games, for the construction and knowledge elicitation task. In additional to the interactive user interface, games also attract players through its game nature, puzzle design, and story attractiveness. We expect to find that the gaming system not only makes the activity easier for experts, but also allows experts to enjoy the interactive experience of building BNs through game play.

6.3 A Game Approach for BN Elicitation

6.3.1 LinkIT Elicits BN Structure

In current practices, questionnaires/surveys similar to the one shown in Figure 3.1 are very frequently used for the elicitation of BN and other network-like structures. For
example, in [53], the authors use a structured interview, as shown in Figure 6.4, to elicit direct relations between variables in a causal map. Participants are asked to select from a set of options that describe potential relations between variables. Because LinkIT has demonstrated its validity and effectiveness of eliciting relations between (risk) variables from players, we can use it to elicit causal relations and elicit knowledge from domain experts to construct BNs.

Motivated by the causal mapping approach reviewed above, we can use the game to help identify missing links and eliminate indirect as well as bi-directional relations. One potential approach to do this is to present more than two nodes in each LinkIT puzzle. For example, if a puzzle contains three nodes, players are encouraged to provide the most direct relation among these three nodes. In this way, indirect and bi-directional relations are both eliminated. Figure 6.5 shows an example of this application. Players are asked to draw the relations between three nodes, A, B, and C. Because A directly causes B and B directly causes C, there is an indirect relation that A causes C. However, because all three nodes are provided and players are instructed to include direct relations only, the indirect relation can very probably be eliminated in the game play. Similarly, experts are able to identify missing links from such a puzzle format.

To maximize the purposeful gain of each game play, an efficient algorithm for puzzle selection will be useful. It should comprehend the following features.

- It minimizes the number of puzzles that need to be responded by each game player to construct a Bayesian network structure.

- It minimizes experts’ mental load in the game play experience. One aspect of feature is that the algorithm presents puzzles logically and allows players to proceed in a reasonable manner.

- It provides an appropriate challenge level to game players so help them enter the
Iterative computational simulations as well as experimental studies are needed to improve this puzzle selection algorithm.

In summary, given a predefined set of variables of interest (nodes), LinkIT can elicit knowledge from domain experts to build the qualitative structure of BNs. Adaptive puzzle selection algorithms and data analysis algorithms are needed to improve the elicitation process and to build the structure. Figure 6.6 summarizes this process.

**Figure 6.5.** An example application of having three nodes in a *LinkIT* puzzle to eliminate indirect relations flow state [80].

**Figure 6.6.** The process of *LinkIT* eliciting BN structure. The input is a predefined set of variables of interest. Provided by adaptive puzzle selection algorithms and data analysis algorithms, the output is BN structure.
6.3.2 SortIT Elicits Conditional Probabilities

According to the probability elicitation study in Chapter 5, we can utilize the SortIT game and different forms of probability information (such as probability intervals and verbal expressions) to elicit conditional probabilities from domain experts for the construction of BNs. Changing the question prompt can help elicit different probabilities. For example, a question prompt for BN probability elicitation could be “Given that the patient has decease X, what is the probability that he also has decease Y?” As shown in Figure 6.7, provided with a BN structure, the SortIT game elicits data from domain experts to build conditional probability tables for the BN. Again, intelligent puzzle selection and data analysis algorithms are needed to maximize the purposeful gain of the game application.

![Figure 6.7](image)

**Figure 6.7.** The process of SortIT eliciting conditional probabilities for BNs. The input is BN structure along with possible values for each variable in the structure. Provided by adaptive puzzle selection algorithms and data analysis algorithms, the output is the conditional probabilities for the BN structure.

6.3.3 The Proposed Game Approach

Figure 6.6 and 6.7 describe the process of utilizing LinkIT to elicit BN structure and the process of utilizing SortIT to elicit conditional probabilities. Combining them together,
provided with a predefined set of variables of interest and their possible values, we can use the game system to build BNs through the following steps.

- Domain experts play the *LinkIT* game and provide data regarding the relations between variables. The system further runs algorithms to build the acyclic directed network structure for the BN.

- For each variable and its possible values in the network, domain experts play the *SortIT* game to estimate the conditional probability distributions.

- The system uses the game play data to derive conditional probability distribution tables for the BN.

Figure 6.8 summarizes this process.

![Diagram](image)

**Figure 6.8.** The process of the game system eliciting Bayesian networks. The system requires an input of a predefined set of variables of interest and their possible values.

### 6.3.4 *NodeIT*: Eliciting Nodes from the Game Play?

Although the game does not allow participants to communicate with each other in the game play, experts can view each other’s responses to the puzzles at the debriefing stage (Section 2.1.4). In this way, we believe that an agreement between experts will be
reached after several rounds of game play. Furthermore, the debriefing feature provides an opportunity of eliciting nodes from experts. At the end of each game, we could ask players “Are there anything missing in the games?” and allow them to enter new variables to the database.
Bibliography


YAN CAO

EDUCATION
Ph. D. College of Information Sciences and Technology, Penn State University 2014
M.S. Department of Physics, Penn State University 2009
B.S. Department of Physics, University of Science and Technology of China 2007

RESEARCH EXPERIENCE
USING GAMES FOR KNOWLEDGE ELICITATION – 2011-2013
Ludic elicitation - Penn State University
Developed two browser-based casual games, LinkIT and SortIT; Explored the validity and performance of a game based approach for relation elicitation, ranks elicitation, and probability elicitation; Conducted experimental studies to measure the validity of elicited data from the game based approach and the user experience in the activities

CROWDSOURCING FOR DATA COLLECTION - 2011-2012
Participatory risk management - Penn State University: Developed an iPhone app for participatory risk data collection, verification, and risk management and conducted campaigns to have university students and citizens to participate
Participatory aurora data collection and forecasting - Los Alamos National Laboratory: Built a website for citizens to share aurora event coverage and forecasting; supported visualization on Google maps, and incorporated twitter data related to aurora sights

TEACHING EXPERIENCE
Graduate teaching assistant for a number of undergraduate courses in Information Sciences and Technology, Security and Risk Analysis, and Physics 2007-2009

SKILLS
Data: Data mining, data modeling, relational database design and management
HCI: Game and serious game design, artificial intelligence, risk analysis, user experience

Address: 1601 Willow Road, Menlo Park, CA 95134