AFFECTIVE ENGAGEMENT IN INFORMATION VISUALIZATION

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Ya-Hsin Hung

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2019

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Steven Landry

School of Industrial Engineering

Dr. Paul Parsons

Department of Computer Graphics Technology

Dr. Robert Proctor

Department of Psychological Sciences

Dr. Anne Traynor

College of Education

Approved by:

Dr. Steven Landry

Acting Head of Industrial Engineering

ACKNOWLEDGMENTS

First, I would like to express my deepest appreciation to my two co-advisors. I'm extremely grateful to Dr. Paul Parsons for his patience, dedication, expertise, and mentorship that helped me throughout the process of my research and writing of this dissertation. The completion of my dissertation would not have been possible without his professional and personal support. I would also like to extend my deepest gratitude to Dr. Steven Landry. He guided me during the toughest times in my Ph.D. pursuit, and taught me how to overcome various challenges with self-discipline and determination.

I'm very grateful to have Dr. Robert Proctor and Dr. Anne Traynor as my committee members. They provided invaluable insight into the human factors and instrument development components of this work, and always shared their constructive advice on the design and execution of my studies. I very much appreciate the opportunities to learn from them. I am also thankful to my former advisor Dr. Ji Soo Yi, who led me to the world of information visualization, and for being my advisor in the earlier years of my Ph.D.

My lab mates in the DVC lab have contributed enormously to my personal and professional time at Purdue. They are not only my academic collaborators, but also my friends. I'd also like to acknowledge the assistance of the instructors and tutors at the Purdue Writing Lab throughout the development of my dissertation. Additionally, I had the pleasure of working with past members of the HIVE Lab.

My time at Purdue was enjoyable because of the many friends and groups that became a part of my life. My student life here was enriched by the Human Factors and Ergonomics Society - Purdue Chapter, Industrial Engineering Graduate Student Organization, and I Love Taiwan Club.

I would like to give special thanks to my MA advisor Dr. Chieh-Hsu Chen at National Cheng Kung University, who gave support and played a significant role in my decision of studying my Ph.D. in the United States.

Lastly, words cannot express how grateful I am to my family. I would like to say thank you to my grandmother and my late grandfather, although away from them is difficult, I appreciate them for letting me pursue my dreams. I would like to thank my brother as well, for taking care of everyone when his sister was not around. And most of all, for my beloved parents, who love me and support me; their encouragement was what sustained me thus far.

Thank you.

Ya-Hsin Hung

August 2019

PREFACE

All of the work presented hence forth was conducted in the DVC Lab (Design, Visualization, & Cognition Laboratory) at Purdue University, West Lafayette. All projects and associated methods were approved by Purdue University's Human Research Protection Program [IRB protocol#: 1611018468, 1703018955, 1902021654, 1903021883].

Stage 1 I was the lead investigator and main experimenter of Stage 1 where I was responsible for all major areas of concept formation, data collection, data analysis, as well as the majority of manuscript composition. Ali Baigelenov and Michael Saenz were involved in the qualitative coding process as coders and contributed to manuscript edits. Dr. Paul Parsons was the principle investigator and supervisory author on this study and was involved throughout the study in concept formation, data analysis, and manuscript composition. Note that the work came from a creative process that heavily involved all of them. A preliminary version of Stage 1 has been published [1].

Stage 2 I was the lead investigator of Stage 2, responsible for all major areas of concept formation, item writing, data collection, data analysis, as well as manuscript composition. Dr. Anne Traynor advised the research methods of instrument development. Dr. Paul Parsons was the principle investigator and supervisory author on this study and was involved throughout the study in concept formation and study administration. Note that the work came from a creative process that heavily involved all of them. A preliminary version of Stage 2 and part of literature review has been published [2].

Stage 3 I was the lead investigator and main experimenter of Stage 3 where I was responsible for all major areas of concept formation, data collection, data analysis, as well as the majority of manuscript composition. Dr. Anne Traynor instructed data collection and data analysis throughout the study. Ali Baigelenov contributed to experiment preparation and manuscript edits. Dr. Paul Parsons was the principle investigator as well as supervisory author on this study and was involved throughout the study in concept formation and manuscript composition. Dr. Steven Landry supported and reimbursed expense for the experiments. Note that the work came from a creative process that heavily involved all of them. A preliminary version of Stage 2, Stage 3, and part of literature review has been published [3].

Stage 4 I was the lead investigator and main experimenter of Stage 4, responsible for all major areas of concept formation, data collection, data analysis, as well as manuscript composition. Dr. Anne Traynor instructed data collection and data analysis throughout the study in concept formation and manuscript edits. Dr. Paul Parsons was the principle investigator as well as supervisory author on this study and was involved throughout the study. Dr. Steven Landry supported and reimbursed expense for the experiments.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

Figure Page

ABSTRACT

Hung, Ya-Hsin PhD, Purdue University, August 2019. Affective Engagement in Information Visualization. Major Professor: Steven Landry and Paul Parsons.

Evaluating the "success" of an information visualization (InfoVis) where its main purpose is communication or presentation is challenging. Within metrics that go beyond traditional analysis- and performance-oriented approaches, one construct that has received attention in recent years is "user engagement". In this research, I propose Affective Engagement (AE)– user's engagement in emotional aspects as a metric for InfoVis evaluation. I developed and evaluated a self-report measurement tool named AEVis that can quantify a user's level of AE while using an InfoVis. Following a systematic process of evidence-centered design, each activity during instrument development contributed specific evidence to support the validity of interpretations of scores from the instrument. Four stages were established for the development: In stage 1, I examined the role and characteristics of AE in evaluating information visualization through an exploratory qualitative study, from which 11 indicators of AE were proposed: Fluidity, Enthusiasm, Curiosity, Discovery, Clarity, Storytelling, Creativity, Entertainment, Untroubling, Captivation, and Pleasing; In stage 2, I developed an item bank comprising various candidate items for assessing a user's level of AE, and assembled the first version of survey instrument through target population and domain experts' feedback; In stage 3, I conducted three field tests for instrument revisions. Three analytical methods were applied during this process: Item Analysis, Factor Analysis (FA), and Item Response Theory (IRT); In stage 4, a follow-up field test study was conducted to investigate the external relations between constructs in AEVis and other existing instruments. The results of the four stages support the validity and reliability of the developed instrument, including: In stage 1, user's AE

characteristics elicited from the observations support the theoretical background of the test content; In stage 2, the feedback and review from target users and domain experts provides validity evidence for the test content of the instrument in the context of InfoVis; In stage 3, results from Exploratory and Confirmatory FA, as well as IRT methods reveal evidence for the internal structure of the instrument; In stage 4, the correlations between total scores and sub-scores of AEVis and other existing instruments provide external relation evidence of score interpretations. Using this instrument, visualization researchers and designers can evaluate non-performance-related aspects of their work efficiently and without specific domain knowledge. The utilities and implications of AE can be investigated as well. In the future, this research may provide foundation for expanding the theoretical basis of engagement in the fields of human-computer interaction and information visualization.

Keyword: information visualization, user experience, instrument development, evaluation

1. INTRODUCTION

Emotions have a crucial role in the human ability to understand the world, and significantly influence how user experience is shaped [4, 5]. Emotions are also well-known to be essential to rational behavior and decision making [6]. Norman [7] proposed the concept of the emotional system, and how its three levels influence one another to create our overall emotional experience of the world.

Engagement is another important aspect of user experience, and has been a popular topic in HCI (Human-Computer Interaction) domain [8]. However, currently people's opinions upon the definition as well scope of user engagement in infoVis (Information Visualization) are rather mixed and unclear [9–11].

Furthermore, little research has focused on developing guidelines and instruments for assessing user engagement in infoVis. While there are various attempts to assess or to measure this phenomenon (e.g., [12–14]), little of them focus on the underlying data—which is a core component of information visualizations. Moreover, none of the existing related studies mainly contributes to the emotional aspects.

Therefore, this dissertation studied the gap between "user engagement" and "information visualization", and focused on the "emotional" aspect—which is **Affective Engagement** (AE) in information visualization. Theoretical assumptions would become useful when they can be tested empirically [15]. Hence, to make AE be testable, a proper measurement tool should be developed and used.

Therefore, the aim of this dissertation was to:

1. Understand AE and its scope in infoVis domain;
2. Investigate the characteristics of AE in infoVis domain;
3. Develop an AE measurement in infoVis domain; and

4. Evaluate the developed measurement for its validity and reliability.

1.1 Survey Instrument for Affective Engagement

The work is not aimed at assessing long-term emotional investment of users—e.g., in situations where a visualization tool is being used everyday in a work setting. Rather, this is aimed at assessing emotional investment of short-term uses of visualizations. Examples include viewing or interacting with visualizations in interactive news stories, public information displays, museums, and interactive textbooks.

1.2 Target Audience

The target audience for this research is people looking for quick and easy ways to evaluate AE. While this target clearly fits visualization practitioners, academics and other researchers often have need for quick and easy evaluation methods, too. Practitioners often face constraints, such as time, money, and other equipment limitations, that make lab-based user testing not feasible. Evaluation methods involving specialized equipment (e.g., eye trackers, EEGs) or considerable money and space to run user studies—while certainly valuable—are outside the scope of our concern here.

Practitioners can benefit from quick and easy evaluation methods that can still provide actionable information regarding AE. When I say we want evaluation to be "quick and easy", I want all stages (i.e., conducting, analyzing, interpreting) of the process to be both quick and easy.

- By **Quick**, we mean a minimal time spent to conduct the testing, to analyze the collected data, and to make sense of the results for making further decisions.
- By **Easy**, we mean there is no need for specialized domain knowledge to conduct the testing, no need for specialized equipment to collect the data, and the collected data is easy to process and easy to interpret.

With the above criteria, I believe that *a concise self-report survey instrument that can quantify AE* can be an appropriate tool for visualization designers wanting to evaluate AE for communicative purposes.

1.3 Development of a Survey Instrument

A self-report measurement tool such as survey instrument that can quantify such construct in InfoVis domain should be created. Therefore, in the rest of this dissertation, I developed and evaluated a survey instrument named "**Affective Engagement Visualization Survey**" (AEVis) that can quantify a user's level of AE while using an InfoVis.

Employing an evidence-centered design approach [16], in every activity during instrument development contributes specific evidence to support the validity of later interpretations of scores from the instrument.

Four stages were established for the development:

1. In **Stage 1**, the role and characteristics of AE was exmained in evaluating information visualization through an exploratory qualitative study to elicit indicators of AE;
2. In **Stage 2**, an item bank comprising various candidate items for assessing a users level of AE was developed, several tryout sessions and domain expert review were conducted target population, and a draft of survey instrument for AE then was proposed;
3. In **Stage 3**, the developed survey instrument was tested in the field test studies. Three analytical methods were applied during this process: Item Analysis, Factor Analysis, and Item Response Theory; and
4. In **Stage 4**, a follow-up field test study is established to investigate the external correlations between constructs in AEVis and other existing instruments.

2. LITERATURE REVIEW

2.1 Overview

In this section, the investigation of related studies would mainly focus on *User Engagement* in the areas of HCI and interactive technology. First, the definitions, the characteristics, and the measurements of user engagement in interactive technology were reviewed. Due to the fact that interactive visualizations nowadays are also considered as one kind of digital system, studies of user engagement in technology will shed some light on determining user engagement in InfoVis domain. Finally, the role and significance of user engagement in InfoVis, and discuss challenges in its characterization and assessment, would be explored and discussed.

2.1.1 What is User Engagement?

Because engagement is of interest to researchers from many disciplines, each having its own priorities and concerns, it is likely impossible to reach an all-inclusive definition of user engagement. While seemingly an important aspect of user experience, user engagement lacks a clear, agreed-upon definition in general [8]. Engagement is a major theme of research within HCI and other related fields, the need to understand users' experiences has motivated a focus on user engagement across multiple area of research. A meta-analysis on user engagement in HCI domain have been conducted by Doherty [8].

On the other hand, a small number of researchers have recently begun to explore the notion of engagement in InfoVis. Mahyar et al. [9] argued for viewing engagement at multiple levels, and have proposed a preliminary taxonomy for evaluating user engagement based on Bloom's taxonomy [17, 18], which is a three hierarchical models

used to classify educational learning objectives into levels of complexity and specificity [19]. To study the effect of animation and pictographs in data videos (a type of animated visualization), Amini et al [20] developed an engagement questionnaire for assessing engagement for different types of visualizations. Also, Windhager [21] tried to explore the design space of engaging climate change visualizations for public audiences, and specified that the concept of engagement combines the cognitive and affective effects of InfoVis, together with the user engagement previously defined by O'Brien [22].

Saket et al. [23] argue for going beyond usability and performance in InfoVis evaluation, and have provided an overview of recent work related to user experience. Others have examined engagement indirectly—e.g., looking at whether storytelling in InfoVis engages users [24], and how aesthetic concerns engage users [25]. At this point, research on engagement in InfoVis is still in its infancy. Definitions and assessments have largely been borrowed and adapted from other disciplines.

Although scholarship on engagement exists in other technology related fields (e.g., [26–29]), InfoVis deserves its own treatment due to its specific characteristics not necessarily present or prevalent in other disciplines—e.g., the abstract nature of data and information, visual encoding and representation, cognitive and perceptual issues, and interaction.

In InfoVis, a focus on non-utilitarian uses of visualizations has been increasing in recent years. Similarly there is an increasing interest in visualizations used for communication and presentation purposes rather than analysis ones [30]. Since traditional performance-oriented approaches (e.g., measuring time spent, calculating error rate) can only be useful for performance metrics such as task accuracy, those approaches rarely work well for affective-relative metrics such as perceived satisfaction. Therefore, for visualizations that are mostly used in communication or presentation purposes, a hedonistic perspective of user experience could be more useful than a utilitarian one.

2.1.2 Characterizing User Engagement

Engagement is a complex construct—it is abstract, not directly observable, and composed of multiple parts. Complex constructs can be difficult to define, as they cannot be directly accessed and can be measured only via an observable phenomenon in which they are manifest [31].

User engagement can be identified in multiple disciplines including psychology, education, games, and HCI, all of which characterize engagement differently. For instance, from a psychology perspective, engagement is often discussed in relation to flow, positive psychology, fulfilment, and motivation (e.g., [32,33]). In education, the concept of student engagement has received much attention, and is usually discussed in terms of motivation, achievement, and interpersonal relationships (e.g., [34, 35]). Within the context of gaming, engagement is believed to be a generic indicator of game involvement [13], and some researchers think an engaging experience is encouraged by the sensory appeal of the system and the level of feedback and challenges a user receives from the system [12, 26].

In HCI, several theoretical frameworks have been proposed [22]. User engagement has been viewed in the context of flow and fluid interaction, leading to satisfying and pleasurable emotions related to curiosity, surprise, and joy [27]. It has also been defined as the emotional, cognitive, and behavioral connection that exists between a user and a resource in time or possibly over time [29]. User engagement is also believed to be the positive interaction quality of the user experience, and has been associated with being captivated and motivated to use a website [36]. Sometimes it is treated as user's level of involvement with a product [37]. Additionally, terms such as flow [38, 39], immersion [26, 40], enjoyment [41], and playfulness [42–44] have been mentioned in related research areas, some of which are close to the concept of user engagement.

As O'Brien [45] notes, the scope of engagement must be determined before constructing a useful definition. Inspired by Lucero et al.'s work [26], we reviewed ap-

proximately 150 papers in related disciplines such as website analysis, game design, education, psychology, and HCI, and compiled a list of potentially relevant characteristics. In the end, 11 engagement characteristics were selected that had the highest frequency in the literature and were most relevant to engagement in InfoVis: Aesthetics, Captivation, Challenge, Control, Discovery, Exploration, Creativity, Attention, Interest, Novelty, and Autotelism [2].

2.1.3 Affect as Transient Emotion

To investigate the emotional aspects of engagement, it is reasonable to also examine characteristics of human emotions and how these characteristics might influence engagement as a phenomenon in the context of InfoVis. When studying emotion, there is a continuum of timeframes that can be investigated, from transient emotional states, to longer-term mood states, to long-term traits or dispositions [46]. Short-term emotional states are relatively brief episodes with clear onset and offset, whereas moods persist over longer timeframes and do not fluctuate as much [47]. Transitory emotional states are more likely than long-term moods to be related to particular events or stimuli [48]. Thus, it is reasonable to focus on short-term states when investigating emotional impact or involvement when users are viewing or interacting with visualizations.

2.2 Assessment of User Engagement

For evaluating user experience, a trend towards non-utilitarian concerns has been emerged in the past decade [5, 6]. Similar trend can be found in HCI as well. While the second wave HCI focus on groups working with a collection of applications, where situated action, distributed cognition, and activity theory were important sources of theoretical reflection [49]. In the third wave of HCI, the use context and application types are broadened, and intermixed; the focus of the third wave seems to be defined in terms of what the second wave is not: non-work, non-purposeful, non-rational,

etc [50, 51]. Scope of cognition has expanded to emotion and design as well [52]. Thus, in the following sections, various types of assessment techniques and methods would be reviewed for a better understanding of the bigger context.

2.2.1 Self-Report Methods

Self-report instruments is the most widely used type of physical activity measure which refers to the methods that rely on what users say or recall about their experience [53,54]. Advantages of self-reporting methods include interpretability (easy to interpret the date), information richness (great quantity and breath of the information), and practicality (cost-efficient and inexpensive) [55]. On the other hand, the main disadvantage for self-assessments is credibility (e.g., social desirability bias [56, 57] and measurement bias [58]), as self-reports are subject to various inaccuracies.

Interview

Interview is one of the most common and powerful way to understand people and collect qualitative data [59]. It can be conducted as face-to-face exchange between two people, with groups of people, or even through other media such a telephone, tele-conference applications etc. In terms of the format of interview, interview can be categorized as structured, semi-structured, and unstructured [60]. While structured interview utilizes closed-ended questions, unstructured interview tends to use open-ended and in-depth questions, semi-structured interview, on the other hand, combines the above two. In general, the interview would be recorded (voice or video) and be transcribed into text format, then be analyzed qualitatively [61]. In O'Brien's study on defining user engagement, she employed semi-structured interview with critical incident techniques, and then identified the attributes of user engagement in technology [62]. Recently, there are also researchers who have employed interview method for assessing engagement or other related constructs in various environment settings [63–66].

Verbal Protocol

Verbal protocol is one of the primary tools to evaluate usability in the HCI domain [67, 68]. The main idea of think aloud and think after approaches is to ask people verbalize their thoughts and feelings, this approach can be used to observe insight into participant's cognitive processes [69]. While think aloud approach refers to do verbalization during the task or activity, think after refers to do it after the session (retrospective). Similar to data collection in interview, researchers usually record participants' voice and transcribe it into textual format.

However, when utilizing think aloud protocol, there are some limitations. For example, Ericsson and Simons believed that verbalizing participants' thought processes does not change the sequence of thoughts, and therefore their task performance should not be changed as a result of thinking aloud [70]. Still, some studies found that additional cognitive activities are required in order to produce the overt verbalization of the thoughts; these activities therefore generate negative effect (e.g., decreases task performance, produces biased accounts of the thoughts) on the task performance [71]. Therefore, there are many other updated verbal communication approaches being proposed [72]. Finally, since users usually think faster than they can speak, their thoughts are expected to be much more complex than they can verbalize [73]. Since think aloud protocol can be easily adapted to different contexts or settings, we can see that some researchers applied this method when they were interested in assessing users' engagement of their works (e.g., [74, 75]).

Questionnaires

Questionnaire is a popular data collection method, and is sometimes referring to survey or instrument. Since it is easy to conduct and because of its low-cost, questionnaire methods are very popular in many domains. There are many types of questionnaires, they can be paper-based or electronic; they can be used on an

individual or on a group of people; they can be designed as close-ended (quantitative) or open-ended (qualitative).

The game research community has been developing the concept of gamer engagement for a long time, even though the terms they used are slightly different from one to another. In Jennett et al.'s study [76, 77], they developed and validated an immersive questionnaire called Immersive Experience Questionnaire (IEQ), this questionnaire is used to measure the immersion that players experienced in a single scale that varies from low levels to high. Another well-established measurement for gaming experience is Gaming Engagement Questionnaire (GEQ). It measures the level of engagement that player experienced in video game-playing in order to evaluate the influence of violent video games on the players [13]. Finally, another group of researchers proposed an engagement metric for web pages called the User Engagement Scale (UES) [36].

There are other questionnaires that are related with user engagement in a higher level. For instance, refer to the Flow concept from Csikszentmihalyi, Flow Questionnaire (FQ) is a set of standardized questionnaire that can be used to measure user's level of flow [78], and Experience Sampling Method (ESM) is another questionnaires for a similar purpose [39]. For assessing presence, there are some instruments as well [79]. Additionally, Playful Experiences (PLEX) is a framework that categorizes playful experiences and can be utilized as an evaluation tool for artifacts such as tangible digital game or interaction designs [42, 80].

Finally, in the area of infoVis studies, several attempts have been made to develop proper survey instruments for evaluating users' various types of engagement [2,11,46].

2.2.2 Physiological Measurements

The main idea of physiological measurements on user engagement is to look for the relationship between physiological processes and thoughts, emotions [81], and behavior. Compared to self-report approach, they can produce more objective data.

However, constructs such as engagement cannot be quantified easily. Although various physiological indicators can be measured directly, such as blood pressure [82], heart rate [83], nervous system activity in general [84], and so on, these methods can be considerably costly and lengthy, limiting their scalability and practicality. Furthermore, they require considerable interpretation to make causal connections to subjective phenomena [85].

2.2.3 Behaviour Measurement

Eye-Tracking

Eye-tracking, as a popular technique in studying usability on artifacts or technology, is aimed to identify and analyze patterns of visual attention of individuals as they perform specific tasks (e.g. reading, searching, and scanning a document etc.). Measures such as pupil dilation, gaze fixation, and visit count can be used as indicators of task difficulty, fatigue, mental activity, and intense emotion [86, 87]. To conduct eye tracking studies, an eye tracker is necessary. It is a device that includes infrared projections that can illuminate a user's cornea (it can be bright or dark pupil eye tracking), and then a set of infrared cameras would gather the reflection patterns and its positions to calculate the pupil locations over time [88]. Finally, after the task was done, the locations of gaze points would be mapped and overlaid on the recordings for further analysis.

Eye fixation [89] and saccadic eye movement [90, 91] are believed to be related with the allocation of visual attention. There is considerable amount of eye tracking research papers study the relationship between eye movement and visual search [92, 93]. There are also studies trying to find out the implications of eye movement such as fixation and visit on user engagement of reading tasks [94]. Finally, in terms of pupil size, pupillary dilation is related with human cognitive workload [95–97], memory [98, 99], and emotional arousal [100, 101].

However, there are several limitations on eye tracking methods. First, eye tracker is an additional device that might influence participants' natural behavior [102]. Second, attention or user preference is not completely equal to user engagement. Tthey might have overlap with each other, but they can't be the only indicator of user engagement [26].

Cursor Tracking

Cursor tracking (or mouse tracking) refers to the movement of a user's mouse cursor on the interface or entire screen. This method usually utilizes additional software or JavaScript code (for websites) to collect the x, y coordinate data (on the screen or in a web page) over time. The metrics in cursor tracking can be clicks, visit counts, total distance traveling, movement speed, scroll speed, and frequency etc. [103]. Mouse movement has correlations between eye movement and gaze [104, 105], and in particular is related with attention [106] and preferences [107].

Web Analytics

While self-report method and physiological measures are usually applied to a small group of people, researchers in the HCI domain proposed a considerable amount of techniques on measuring user engagement in a larger scale. For example, Google proposed a metric framework to measure the level of user engagement of web page called HEART (Happiness, Engagement, Adoption, Retention, and Task Success) [37]. On the other hand, Dupret and Lalmas tried to assess user's depth of engagement on websites with absence time and clicks [108]. Targeting on-line shopping environments, O'Brien and Cairns have evaluated User Engagement Scale (UES) in an on-line platform [109]. Another attempt from Thomas and O'Brien was to measure user engagement of on-line forms with UES as well [110]. Finally, Attfield et al. proposed engagement characteristics among three dimensions that are associated with user engagement: emotional, cognitive and behavioral [29].

2.3 Existing Survey Instruments

To provide context before presenting my own work, I had to investigate established self-report instruments that researchers and practitioners can use to evaluate their visualizations. To do so, I conducted a brief survey of relevant evaluation instruments. There are two intentions here: (1) a collection of these instruments can be a valuable resource on its own; and (2) the survey helped to highlight where gaps might be—for communicative issues in general, and AE in particular – in the context of InfoVis.

2.3.1 Method

To conduct the survey, I searched for and collected relevant self-report instruments. The initial search was very broad; besides some general resources from HCI and UX (User Experience) handbooks [111–113], I also searched online using the following keywords: "visualization", "user experience", "engagement", "communication", "persuasion", "emotion", "survey", "questionnaire", "scale" and their various combinations. Although I found many instruments related to communication, affect, satisfaction, and various psychological constructs, I excluded all that were not concerned with human-technology relationships. Thus we excluded instruments dealing with constructs such as human-human communication, anxiety, customer satisfaction, and so on.

In the end, I settled on 3 inclusion criteria; Eeach instrument should:

1. Be concerned with human-technology relationships;
2. Be associated with a publication; and
3. Not require specialized equipment.

Several internal research team meetings were to identify characteristics that are relevant for communicative visualization and could be used to code the instruments. The characteristics include whether an instrument is concerned with:

(a) communicative effectiveness of the technology;

(b) visual aspects of the technology;

(c) performance (e.g., time, error);

(d) user engagement;

(e) affect;

(f) a particular platform or scenario; and

(g) whether the instrument has a commercial version that needs to be purchased.

Examples of the characteristics include: communicative effectiveness: "*Prompts for input is confusing/clear.*" (R3 [114]); visual aspects: "*The screen layout of this website is visually pleasing.*" (R22 [12]); performance metrics: "*I can recover from mistakes easily and quickly.*" (R12 [115]); engagement metrics: "*I really get into the game.*" (R20 [13]); affect metrics: "*The system is somewhat intimidating to me.*" (R13 [116]).

The review is not meant to be exhaustive, yet due to the systematic approach, I believe it is reasonably representative of a more complete sample. By following two strategies—investigating popular books and conducting our own search—we believe that we have covered at least the popular and well-established instruments.

2.3.2 Findings

The survey resulted in 24 instruments that met the inclusion criteria. A summary of these is shown in Table 2.1. For each instrument, I list the name, publication year, the construct being evaluated by the instrument, the total number of items (questions) or heuristics included, and the instrument characteristics described previously.

In general, most of the collected instruments deal with constructs like usability and user experience, and more than half are developed based on general system/technology/artifact platforms. Usability oriented instruments usually have substantial numbers of performance-related items and few items related to affect or engagement (e.g., R3 [114], R8 [123]). On the other hand, engagement-related in-

Table 2.1.
24 survey instruments [12, 13, 76, 114–133] meeting all inclusion criteria, ordered chronologically. Rows are collected survey instruments, columns are characteristics of the survey instruments (● = "mostly satisfied", ▲ = "partially satisfied", and ✓ = "Yes").

	Name	Year	Main Construct	# of items	Communicative effectiveness	Visual aspects	Performance metrics	Engagement metrics	Affect metrics	Commercial version	Specific Platform/Scenario	Intended platform
R1	System usability Scale (SUS)	1986	Usability	10			●		▲			System/Technology
R2	NASA Task Load Index (TLX)	1986	Subjective workload	6			▲		▲			Interface/System
R3	Questionnaire for User Interface Satisfaction (QUIS)	1988	Satisfaction	27	▲	▲	●	▲	▲	✓	✓	Interface
R4	Perceived Usefulness and Ease of Use (PUEU)	1989	Usefulness and Ease of Use	12			●				✓	System/Technology
R5	The After-Scenario Questionnaire (ASQ)	1990	Satisfaction	4			●					System/Technology
R6	The Post-Study System Usability Questionnaire (PSSUQ)	1992	Usability	16	▲	▲	▲	▲	▲			System/Technology
R7	Nielsen's Heuristic Evaluation	1994	Usability	10	▲	▲	●	▲	▲			System/Technology
R8	Computer System Usability Questionnaire (CSUQ)	1995	Usability	19	▲		●		▲			System/Technology
R9	Software Usability Measurement Inventory (SUMI)	1995	User experience	50	▲		●	▲	▲	✓	✓	Software
R10	Presence Questionnaire Item Stems	1998	Presence	32			▲	●	▲		✓	Virtual environment
R11	Website Analysis and Measurement Inventory (WAMMI) Questionnaire	1998	User experience	20	▲		▲	▲	▲	✓	✓	Website
R12	USE Questionnaire	2001	Usability	30			●	▲	▲			System
R13	Unified Theory on Acceptance and Use of Technology (UTAUT)	2003	Technology Acceptance	31			▲	▲	●			Technology
R14	Fun questionnaire	2005	User experience (fun)	14			▲	●	●			Educational system
R15	Cognitive absorption and TAM	2005	TAM+cognitive absorption	22			▲	▲	▲		✓	System/Technology
R16	The Single Ease Question (SEQ)	2006	Ease of use	1			●				✓	System/Technology
R17	Immersive Experience Questionnaire (IEQ)-a	2008	Experience of immersion	31	▲	▲		●	●		✓	Video game
R18	Immersive Experience Questionnaire (IEQ)-b	2008	Experience of immersion	33			▲	●	●		✓	Video game
R19	User Experience Questionnaire (UEQ)	2008	User experience	26	▲	▲	▲	▲	▲	✓		System/Technology
R20	Gaming Engagement Questionnaire (GEQ)	2009	Deep engagement	19				●	●		✓	Video game
R21	The Subjective Mental Effort Question (SMEQ)	2009	Ease of use	1			●				✓	System/Technology
R22	User Engagement Scale (UES)	2010	User engagement	31		▲	▲	●	●			System/Technology
R23	Measurement Model of User Engagement	2015	User engagement	12			▲	●	●		✓	Website
R24	Standardized User Experience Percentile Rank Questionnaire (SUPR-Q)	2015	User experience	8			▲	▲	▲	✓		System/Technology

struments have more focus on affect and engagement, with fewer performance-related items. (e.g., R20 [13], R22 [12])

The contexts in which presentation or communicative visualizations are used are different from many of the instruments in Table 2.1 (e.g., to influence or to persuade viewers). Thus, while the compiled instruments may be a useful resource for evaluating communicative and narrative visualizations, the existing instruments are not entirely suitable for the following reasons:

- **Scope:** Most surveys that include affect- or engagement-related items aim to cover a much broader construct. Thus the relevant information gained about AE may not be very substantial (e.g., may be related to only 1 or 2 items). Also, there are problems with using only portions of an instrument without using it in its full and originally intended context [].
- **Specific media or environment:** Some instruments are measuring a construct that is tied to a specific medium or context of use that is not very relevant for communicative visualization (e.g., video game [13], presence in virtual environment [125]).
- **Context of measurement scale:** Although some instruments share similar key factors, the context of their measurement is not always appropriate for communicative visualization. For example, "captivation" may be a sub-component for both "immersion" and "engagement". However, an item that asks "I felt detached from the outside world" (see [76]) is likely not appropriate for communicative and narrative visualizations, yet makes sense for assessing engagement in Virtual Reality.
- **Length:** Some instruments contain a high number of items, and may take substantial time to answer and administer. Although more items may be desirable for precision, long instruments may not be "quick and easy", quickly becoming a barrier for practitioner use.

2.3.3 Need for an Affective Engagement Assessment

From any perspective, user engagement, as with other aspects of subjective experience, is a complex construct that is difficult to define [134]. Recently, as InfoVis investigation has reached beyond usability driven objectives, investigating aspects of subjective experience is increasingly important [2, 9, 10]. It is reasonable to assume that InfoVis researchers and practitioners are interested in how "engaging" a particular visualization is, wanting to measure levels of engagement to predict or determine success [135]. Also, behavior-based metrics (e.g., time spent, see [24]) have previously been employed to quantify "engagement" levels in InfoVis.

Evaluation methods for InfoVis should match the goals of the design situation and the context of use [136]. Because many existing evaluation strategies have been aimed at analysis rather than communication, they are not often suitable for evaluating issues relevant for visualizations for presentation or communication. Thus, there is a need to examine evaluation methods for InfoVis with communication, rather than performance, as the main goal. Which, further confirms the need to development a survey instrument for this purpose.

3. STAGE 1–ELICIT CHARACTERISTICS OF AFFECTIVE ENGAGEMENT

Engagement is a complex construct that has many facets. In this work, in order to develop the theoretical background of affective engagement in Information Visualization (InfoVis), I conducted two mixed-methods studies that involved participants interacting with visualizations to understand characteristics and the development of engagement in the context of InfoVis.

3.1 The Exploration of Emotional Aspect of Engagement in InfoVis

The study started from exploring user engagement when users are interacting with visualizations. I conducted a mixed-methods experiment to investigate our construct of interest in a laboratory setting (IRB protocol#: 1703018955).

Based on literature review [8], I expect "user engagement" is a complex and multiple-dimensions facets construct. Therefore, in order to catch more rich information, both quantitative and qualitative data were collected. There were two primary research questions for the study:

- **RQ1** What are the characteristics of engagement in an InfoVis context?
- **RQ2** What factors contribute to the development of engagement over time as users interact with visualizations?

The goal was to characterize participants' engagement via multiple protocols including think-aloud, eye-tracking, behavioral indicators, and self-assessments. The results of this research can contribute to the development of AEVis as descriptive and explanatory models or frameworks of engagement in InfoVis.

3.1.1 Method

To elicit the components and characteristics of engagement, I asked our participants to interact with interactive visualizations while speaking aloud their cognitive activities as well as feelings. Data from think-aloud, eye-tracking, questionnaires, and semi-structured interviews were collected and analyzed. The lab setting to collect corresponding data is depicted in Figure 3.1. Each participant conducted think-aloud sessions in front of the monitor and input devices (keyboard and mouse) of a desktop PC. A web camera (with microphone) for audio recording was placed on top of the monitor, and an eye-tracker was attached at the lower part.

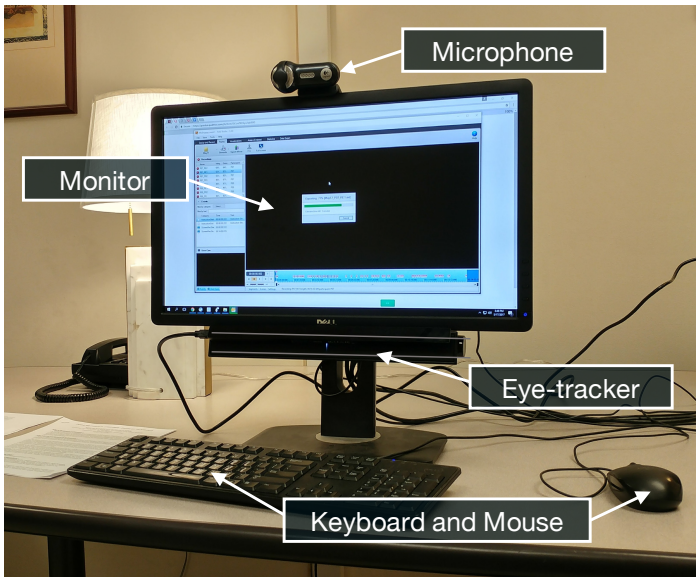


Figure 3.1. Experiment setup of the laboratory study.

In the end, 12 participants were recruited (ages 21 to 47; 8 females, 4 males; all native English speakers). Participants were recruited at Purdue university and their participation was voluntary. A group of this size is generally considered appropriate for studies utilized some labor-intensive qualitative studies (e.g., ground theory) [137].

In this experiment, three methods were utilized: think-aloud [67,69], interviews [60], and questionnaires [59]. The primary data collection method is think-aloud. On top

of that, eye tracking was used to identify and analyze patterns of visual attention as users worked with the visualizations. Specifically, the quantitative data from eye tracking is used to fill in gaps in the qualitative (think-aloud) data, and to triangulate the findings. Mouse tracking was also used to capture users' interaction behaviors and was cross-referenced with eye-tracking data and verbal protocols.

This experiment was focused mainly on qualitative data, requiring intensive analysis procedures that followed a grounded theory approach [138]. Although the bulk of the data analysis was qualitative in nature, I also collected quantitative data, which is primarily be used to triangulate the results. Multiple methods and data were utilized for triangulation, which is a strategy to use more than one approach or data source to accomplish a comprehensive understanding of phenomena [139, 140]. Thus, the approach can be characterized as a *qualitatively driven mixed-methods study* [141]. A mixed-methods approach typically can provide rich data that may not be available from qualitative or quantitative methods alone.

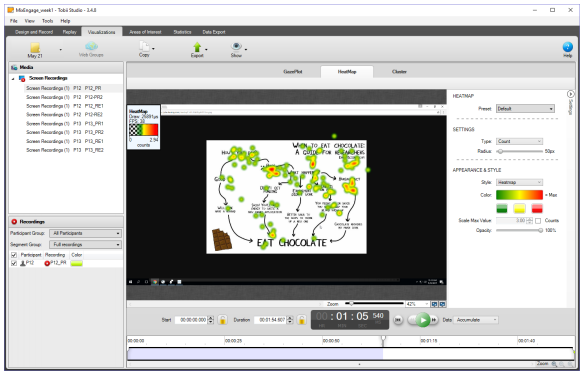


Figure 3.2. Screenshot of Eye-tracking device Tobii studio software.

As the primary data analyzed in stage 1 is the think-aloud data, both eye tracking and mouse tracking were more like the complementary information during qualitative process. Both eye tracking and mouse tracking data were collected via Tobii Studio Professional edition [142], version 3.4.8 (See Figure 3.2), and then exported as video recordings during the analysis.

The collected eye-tracking data and mouse cursor trace are visualized into trace animations and overlaid on the screen recording videos. As shown in the Figure 3.3 A and B, similar to the static gaze plot, the the time sequence of looking or where participants look and when they look there can be shown as animations [143]. The sequence and size of eye-gaze points were illustrated as a series of red circles connected by red lines. The larger circles represent the longer gaze, the size of red circles would increase and decrease as participants moving their focus of attention, and the connected red lines represent the trace of the movement. As for the mouse trace (See B in Figure 3.2), the original cursor would be recorded together with the screen recording, when participant click, a pink circle would appear to signify the mouse action. The steps to overlay mouse and eye tracking traces on the screen recording with user sound in Tobii Studio can be found in the Experiment Material section of Appendix.

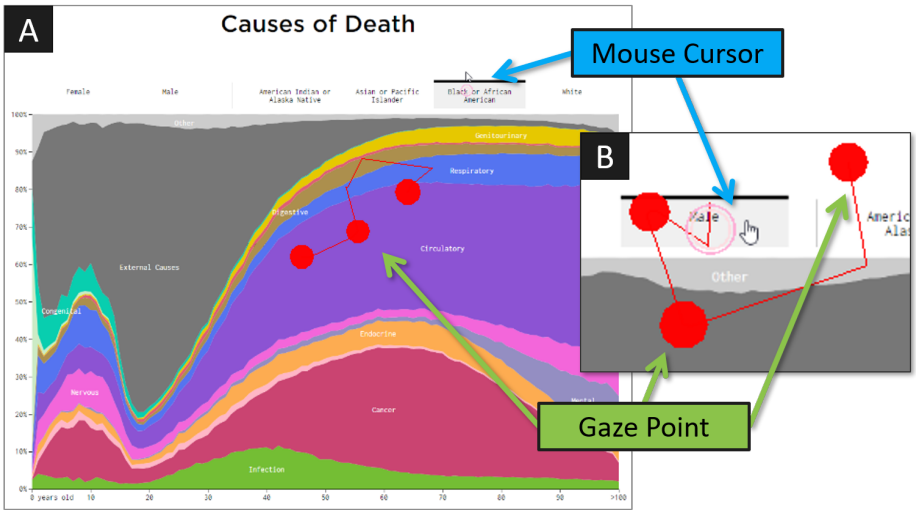


Figure 3.3. Tobii Studio gaze and mouse cursor trace animation.

Finally, Questionnaire is one of the most popular approaches to assessing psychological constructs due to its capability to assess user's subjective feelings and

opinions [2], is also utilized. The example of the survey instrument can be found in the experiment section of Appendix as well.

3.1.2 Experiment

There are various research strategies to be chosen, and many possible methods from which to choose for collecting data on affective engagement. Therefore, below I would articulate the rationales and reasons for the chosen methods for this study.

Think-aloud protocol There are both advantages and disadvantages to using information drawn from think aloud data [144]. Primary advantages of think-aloud method including the ability to provide data about the behavioral, cognitive, and affective processes [145]. As for disadvantages, it is believed that thoughts generated from the long-term memory of subjects are often tainted by perception, participants might incorrectly describe the processes they actually used [146]. Moreover, some participants might have difficulties on the cognitive load of problem solving and speaking [145]. Still, I believe a think-aloud protocol is more beneficial than interviews for the present context. This is primarily because in this study the short time-span in which visualizations are typically used, and high-demanding tasks which will require a lot of high cognitive resources are not expected; issues of think-aloud that mentioned above are not a big concern for this study.

Tested visualizations Due to the impracticality of studying all visualization techniques, only limited numbers and types of tested visualizations can be used in the experiment. Considering the amount of content which will influence both time spent and workload of our participants, after a brief pilot testing with one of the research team members, we decided to have only two tested visualizations for think-aloud sessions.

Although we cannot cover all types of visualizations in the world, it should be a good starting point for studying engagement in different topics and contexts. Hence,

two dimensions were identified to help selecting the visualizations that are reasonably different from each other. To make sure that the two tested visualizations carrying out different types of tasks, we identify following two categories:

- **Explanatory**: Those with pre-defined messages that designers intend to communicate to users (e.g., some journalistic visualizations); and
- **Exploratory**: Those with no pre-defined message that need to be investigated to derive meaning (e.g., visual analytic tools).

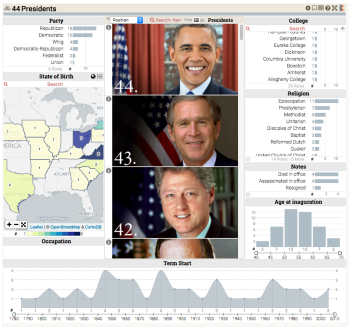
Although there is not always a clear distinction between these two types, it is a useful starting point for studying engagement in different contexts. Additionally, since the interactive features would significantly influence how an user perceive and use a visualization, the research team defined two interaction levels for visualizations:

- **Static**: No interaction; and
- **Interactive**: With interactions or animations. e.g., filtering, zooming, brushing, etc.).

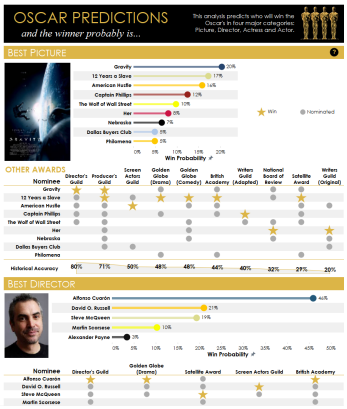
With the two dimensions described above, from several online visualization archives, the "US presidents" [147] (Static, Exploratory) and "Oscar Prediction" [148] (Interactive, Explanatory) were chosen as the two tested visualizations (See Figure 3.4).

Environment Laboratory studies can produce unwanted effects due to the presence of an observer and expectations of completing tasks. In an attempt to mitigate such effects, and to create a more natural and relaxing environment where people will be less aware of that it is an academic study, we employed an incomplete disclosure technique. The study comprised two sessions—the first session involved grading existing visualizations, and the second session involved working with the 2 aforementioned visualizations ("When to eat chocolate" [149] and "Causes of death" [150], see B.3 in the Appendix).

The purpose of the first session was to help participants relax, have more confidence in terms of evaluating visualizations, and have more natural behaviors in the



(a) Exploratory tested visualization "US presidents" [147].



(b) Explanatory tested visualization "Oscar Prediction" [148].

Figure 3.4. Two tested visualization used in the think-aloud session.

second session. Results of the first session had no influence on the second session (participants were not made aware of this until the end of the study).

3.1.3 Data Analysis

In the experiment, data is collected from four main sources: think-aloud, eye-tracking, interviews, and questionnaire. While the think-aloud protocol provided our main data for qualitative analysis, eye-tracking data helped with triangulating verbal protocol data, and was a useful references during qualitative coding. For example, when participant was mumbling or even stop speaking aloud (discussions on this concern is mentioned in [72]), the coders could cross-reference the eye-tracking and mouse tracking data to understand his or her possible intentions, and therefore supplement to the original think-aloud data.

Additionally, scores from the questionnaire responses were considered as references. Specifically, to identify whether participants show any significant pattern across the 11 assessment attributes in the questionnaire. Finally, interview responses

were cross-referenced with the think-aloud results when generating the models of affective engagement.

The audio recordings are transcribed into 24 transcriptions, each participant has two transcriptions for the two tested visualizations. The data analysis was performed in two main phases:

1. A qualitative coding stage, in which a grounded theory approach was employed; and;

2. A triangulation stage, in which data from mixed methods was analyzed and cross-referenced with the findings from the coding stage. Each stage is described below.

Phases 1: Qualitative Coding

In this stage, we selected verbal protocol data from the think-aloud sessions as our primary target of analysis. Following Strauss and Corbin's approach to grounded theory [138, 151], we used three coding phases in our analysis: open, axial, and selective. These phases move progressively from descriptive coding and understanding through to more theoretical analysis and encoding of the data.

Here, *code* refers to the labels generated by the coders and used during the in coding phases to tag excerpts from participants. The coders used a verb + ing style to describe the behaviors or cognitive activities in which participants were engaged during in the transcription. Names of the activities were mostly adapted from the codes, but some of them were modified.

In the **open coding** phase, the coders segmented participants' "raw data" (i.e., transcriptions) into meaningful expressions, then assigned concepts (*codes*) to each expression. The main aim of open coding is to break down the text and assign codes to meaningful chunks. The coders iterated on this phase three times, each time renaming or modifying the codes. During **axial coding**, the aim is to relate codes so that relationships among them emerge. These connections help with identifying

core variables of a phenomenon that can be pursued systematically. In this phase a code co-occurrence chart is used to help identify those interrelationships. During **selective coding**, core variables are selected and used to integrate concepts and categories so that explanations can emerge and theoretical claims can be made. In this phase, the transcripts was re-examinedand; coders selectively sampled new data with the core variables, building a storyline around the core variables and categories to progressively develop two theoretical models of AE.

Three coders continuously had discussions during the coding process to ensure agreement. If there were disagreements, as the primary researcher, I would help to resolve it. Several sessions were conducted to triangulate the coders' styles and strategies.

During each coding phase, the researchers created affinity diagrams to help categorize the codes and to identify similarities and differences among them. Several (5) models were generated and discussed during the coding process. Understanding of the development and structure of affective engagement in InfoVis was gradually generated throughout the process, and finalized at the end of the selective coding stage. Finally, two models grounded in the collected data were proposed: one focusing on the development and fluctuations in AE over time, and the other focusing on the conceptual structure of AE in InfoVis.

To overcome limitations of a think-aloud approach, video recordings—including real time eye-tracking data (eye gaze fixation) and mouse tracking data—were utilized during the coding process. However, note that due to the nature of think-aloud method, sometimes participants forget or are unable to generate verbal responses during the session (e.g., silent moments when participants only stare at a visualization or move the mouse cursor without speaking aloud). In such cases, We utilize eye-tracking and mouse tracking data to supplement the missing information from the recording transcriptions, which will be explained in detail in the following section.

Phase 2: Mixed-Methods Triangulation

In order to capture different dimensions of the same phenomenon, more than one method was utilized in this study. To triangulate the results from stage 1, we utilized participants' responses from questionnaires and interviews as the supplementary of think-aloud data.

After calculating the distribution of codes in each participant's transcriptions, we compared these values with their questionnaire responses, as most of the codes correspond to characteristics in the VisEngage questionnaire (See Experiment Material in Appendix). For example, the *Discovering* activity corresponds to the *Discovery* and *Novelty* characteristics; the *Commenting* activity corresponds to the *Creativity* characteristic; the *Pursuing* activity corresponds to the *Exploration* characteristic; and the *Orienting* activity corresponds to the *Attention* and *Interest* characteristics. For each think-aloud session, the number of activities (determined by coding in stage 1) was cross-referenced with the scores that the participant received from the questionnaire.

3.1.4 Results

Two models of AE emerged as a result of the two-stage analysis. The first is a model of the structural features of AE as users interact with visualizations. The second is a model of the temporal process and fluctuations of AE as it develops over time as users interact with visualizations.

For the following discussions on *activities*, the activities which capitalizing the first letter was derived from the qualitative coding: Acquiring/Obtaining Information, Interpreting, Questioning/Assuming, Pursuing (with purpose), Orienting Attention, Discovering, and Commenting.

The *emotional involvement* will also be discussed by using the following capitalized labels: Positive, Neutral (none), Negative, Surprise, and Curiosity.

All 12 participants were assigned an arbitrary number from 1 to 12. Also, there were two visualizations used in the experiment, one about US Presidents and one about Oscar nominations. Thus, the following format for labeling participant quotes was used: (PNumber, Visualization)—e.g., (P7, Presidents) or (P8, Oscars).

Model 1: Structure and Characteristics of AE

The first model emerged as a result of the qualitative coding process in stage 1. As the research team organized and categorized the codes, they fell naturally into three levels, as shown in Figure 3.5, each level has several associated activities listed within it. The three levels correspond to the following main categories (from low to high):

1. Perception & Action;
2. Understanding and Exploration/Discovery;
3. Emotional Involvement

The perception & action is at the lowest level; two categories (Understanding and exploration/Discovery) are within the middle level since they are highly related; and the emotional Involvement is located at the highest level. Each level has several associated activities listed within it. All three levels are interconnected, where each influences or drives another. Each level is described below.

Perception & Action This is the lowest level in the model. In the perception component of this level, users actively attend to something perceptually within a visualization—e.g., a user locates a visual element within the visualization. In the action component of this level, users perform actions on the visualization. Activities associated with this level are "Acquiring/Obtaining Information" and "Implementing Action."

Acquiring/Obtaining Information Acquiring/Obtaining information refers to an activity where users acquire or retrieve information from the visualization. "In-

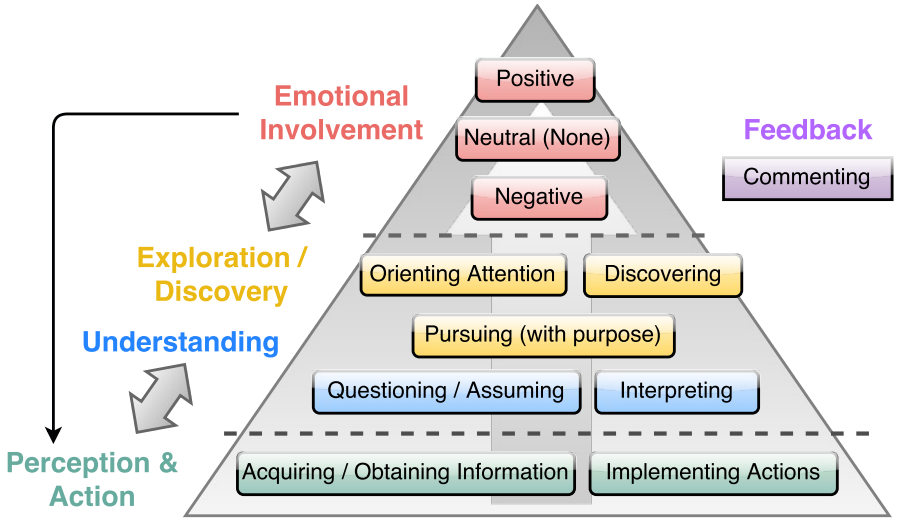


Figure 3.5. Model of the structure and characteristics of AE in InfoVis: There are three levels (from low to high) corresponding to the codes from stage 1 analysis: (1) perception & action; (2) understanding and exploration/discovery; and (3) emotional involvement.

formation" here is broadly construed and refers to all types of visual elements and encodings within the interface. Within this activity, users also describe what they saw in the visualization while they were extracting information. An example of a user Acquiring/Obtaining Information can be seen in the following quote from one participant: *"I can see it says award name: Golden Globe Comedy, Nominee: Amy Adams"* (P2, Oscars).

Implementing Action Implementing Action refers to an activity where users perform an action on the visualization. An example of a user implementing an action can be seen in the following quote: *"I'm going to go ahead and jump down to Amy Adams"* (P12, Oscars).

Understanding Understanding is one part of the middle level in the model. Here, users either perform activities that help them understand or they express their un-

derstanding. Activities associated with this level are "Interpreting" and "Questioning/Assuming."

Interpreting Interpreting refers to an activity where users state, express, or explain their understanding of the content within visualizations. Understandings can be expressed in the form of confirming or rejecting previous hypothesis or assumptions. For example: *"It's interesting how most of the presidents are from the east side of the United States. Most have been Republican."*(P3, Presidents).

Questioning/Assuming Questioning/Assuming refers to an activity where users make an assumption (based on observation, previous knowledge, or a random guess), which can later be evaluated as true or false. For example: *"There's a 50 percent historical accuracy. I'm not really sure what that means, but I guess this is to [sic] this is all predictions [sic]. "*(P4, Presidents).

Exploration/Discovery

This is the other part of the middle level in the model. Here, users are Discovering, Exploring, and Pursuing information due to the visualization or personal thoughts. Activities associated with this level are "Pursuing (with purpose)," "Orientation Attention," and "Discovering."

Pursuing (with purpose) Pursuing (with purpose) refers to an activity where users search for or seek out something specific with a definite purpose. This is in opposition to random exploration or purposeless browsing. For example: *"I'm going to go back and look at other person [sic]who died in office. He just served two years in office."* (P7, Presidents).

Orienting Attention Orienting Attention refers to an activity where users' attention is (re)directed towards something in particular. Several sub-activities are

included within this activity, including but not limited to: "Retrieve prior knowledge" (when a user orients attention towards something held in memory) and "Personal Connection" (getting attached to something or finding something that relates to him- or herself. For example: *"Hugh Stern, great old guy. I can't believe he only has a 17% but he did get the National Board of Review."* (P13, Oscars).

Discovering Discovering refers to an activity where users learn or detect something new while exploring the visualization. The discovered information can be simple facts or complex conceptual knowledge. For example: *"Gravity won two different awards for directors guild and producers guild, and then 12 Years a Slave, well, one, two, three, four, five of them. And then American Hustle won one."* (P13, Oscars).

Feedback

Commenting Commenting refers to an activity where users have a view (opinion/feedback/suggestion) that they come up with after sufficient exploration. They provide references or rationale to what they are talking about. Sometimes users will demonstrate their imaginative or creative ability while commenting. Commenting is not associated with any particular level; rather, it takes place across all three levels. For example: *"it's all the information I would need if I was 10 and I was doing a project on presidents."* (P8, Presidents).

Emotional Involvement

This is the highest level in the model. Within this level, users are feeling or expressing emotions related to the visualization. Activities associated with this level are: "Positive", "Negative", and "Neutral (none)" emotional involvement. In addition, some special case activities are also included such as "surprise" and "curiosity."

Positive Emotional Involvement Users have positive involvement when expressing, including but not limited to: like, impressed and other positive emotions. Users explicitly express positive emotions, but also they can sometimes be inferred from context or tone. For example: *"That's really neat. I really like how when I interact with this, it changes this one over here."* (P2, Presidents).

Neutral (none) Emotional Involvement Users have either no emotional involvement or neutral emotional involvement (i.e., not positive or negative). This includes states such as apathy, lack of care or interest, and others. Similar to positive emotional involvement, neutral emotional involvement can be derived either from explicit statements or context and tone. For example: *"Cool. I don't know. I don't have a strong feeling about this. It's kind of boring to me."* (P8, Presidents).

Negative Emotional Involvement Users have negative emotional involvement, expressing feelings such as dislike and frustration, a sense of loss, struggling and other negative emotions. Similar to the others, negative emotional involvement can be derived either from explicit statements or context and tone. For example: *"I guess the chart at the bottom confuses me, which is why I'm avoiding it like the plague."* (P10, Presidents).

Curiosity Curiosity is seen when users expresses inquisitiveness regarding visualizations, either explicitly (e.g., in the form of a question asked out loud) or implicitly in terms of behavior and tone. For example: *"I wonder if I highlight a dot, it shows all the different facts about them on all the little windows around it"* (P2, Presidents).

Surprise Surprise is seen when an user is surprised by either the visualization itself or the underlying data. For example: *"Wow. If they were assassinated in office, I'm surprised."* (P10, Presidents).

Model 2: Process and development of AE

The second model, as shown in Figure 3.6, consists of the same levels and activities as model 1, but aligned differently. Here, Perception & Action, Exploration/Discovery and Understanding, and Feedback are levels on the horizontal axis (time). Emotional involvement is on a vertical axis, with the middle being neutral (none), above being positive, and below being negative.

In this model, AE is represented using a wave metaphor. Similar to actual waves, users' emotional status fluctuates over time, based on their explorations and understandings of the visualization. Factors such as attention and personal interest were expected to have a considerable effect.

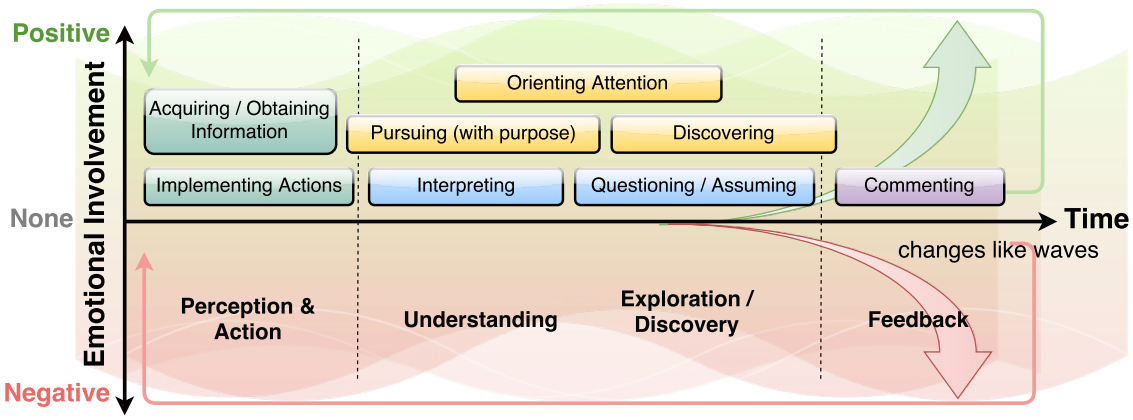


Figure 3.6. Model of process and development of Affective Engagement (AE) in InfoVis: Perception & action, exploration / discovery and understanding, and feedback are levels on the horizontal axis (time). Emotional involvement is on the vertical axis, with the middle being neutral (none), above being positive, and below being negative.

Dimensions of Emotions As described previously, user's emotional status is a dynamic phenomenon that fluctuates back and forth. Affective engagement does not only consist of positive emotions, but negative emotions can also drive the user's

affective engagement. From the data been collected in the experiment, most types of emotions enhance user depth of affective engagement.

Sequence of Activities This model depicts the development of AE over time. A typical temporal process of engagement looks as follows: It starts with perception and action, where users acquire information and act on a visualization. Subsequently, it moves to understanding and exploration/discovery, where users orient their attention, pursue various goals, discover new pieces of information, make interpretations on their perceptions and actions, and ask questions and form assumptions. The emotional involvement level acts like a final state, though it is relevant across the whole sequence. Feedback often follows exploration plus discovery and understanding, although it can occur at any time. Here, engagement can involve both positive and negative emotional involvement, while neutral emotional involvement is equated with disengagement.

3.1.5 Summary

In this section, affective engagement in InfoVis was investigated, with the aim of developing theoretical models of the structure and process of engagement. We employed a mixed-methods approach, relying primarily on qualitative data, yet also cross-referencing qualitative findings with quantitative data. Grounded theory approach was utilized to inductively build two models that are grounded in our experiment data.

Two models of Affective Engagement (AE) in InfoVis were developed using a grounded theory approach for qualitative data and triangulation with qualitative data: **(a) the process and development of AE** (See Figure 3.6), and **(b) the structure and characteristics of AE** (See Figure 3.5). Both models are presented and explained with representative quotes from participants, and the dynamics and interrelationships within the models are elaborated. Yet, considering the survey instrument as an evaluation tool that would be developed based on the results. For situations where people would like to assess visualization user's level of AE, poten-

tial users such as academic researchers and practitioners are expected to be more interested in the positive type of AE. Therefore, in the follow-up study, the research team would focus on the casualties when participants demonstrated positive types of emotional involvement, and elicited factors that could result AE on the positive dimension.

Information visualizations are being increasingly used outside of traditional work contexts. Although InfoVis scholarship has not historically focused on "third wave" HCI concerns [50] (i.e., those including non-utilitarian, non-work contexts, where notions of affect, fun, culture, and others take primacy), such concerns are likely to become more prevalent in the near future. While affect is known to influence user experience and even perceptions of usability, it has not received much theoretical attention in InfoVis.

This study is exploratory in nature, and is not intended to test specific hypotheses or lead to definite conclusions. Rather, it provides an exploration of AE in InfoVis by utilizing ground theory method, and presents inductively derived findings with minimal theoretical assumptions. Specifically, the role of emotion and its characteristics has emerged during the qualitative coding process. Both positive and negative types of engagement existed when users interacting visualizations, and both of them could influence not only participants' reactions, but also how focused they "into" that visualizations.

3.2 Follow-up Study

The previous study contribute to the development of the theoretical background of AE in the context of InfoVis. However, the two models are not operationalized, and the indicators of AE which illustrates the behavior domain of assessing level of AE are not specified in the results. Thus, in this follow-up study, to elicit characteristics of AE in InfoVis, another laboratory study using a mixed-methods (i.e., both quantitative

and qualitative) approach was conducted. We had two primary research questions for the study:

- **RQ1**: What are the characteristics of affective engagement in an InfoVis context? and;
- **RQ2** What factors contribute to the development and fluctuation of affective engagement over time as users interact with visualizations?

3.2.1 Experiment

Similar to the first experiment, The research team recruited 12 participants (ages 21 to 52; 4 females; all native English speakers). Participants were recruited at Purdue university and their participation was voluntary. Again, there were two trials in the experiment. In each one, participants were asked to think aloud while exploring and making sense of a visualization. An eye tracker and a microphone recorded participants' voice, eye movements, and mouse movements. The eye-tracker used in this study was the Tobii X3-120 (See Figure 3.7). After completing two trials, semi-structured interviews were conducted to determine participants prior experience with visualizations and to elicit their thoughts and opinions regarding their own engagement and how and why it developed over time.

This time, the visualizations were selected from the online archive "Information is Beautiful". Although we cannot cover all types of visualizations, we believe it is a useful starting point for studying engagement in different topics and contexts. The assignment of the 24 tested visualizations on 12 participants can be found in Appendix Figure B.5.

3.2.2 Data Analysis

We collected data from three main sources: think-aloud, eye-tracking, and interviews. While the think-aloud protocol provided our main data for qualitative analysis, eye-tracking data helped with triangulating verbal protocol data, and was a useful

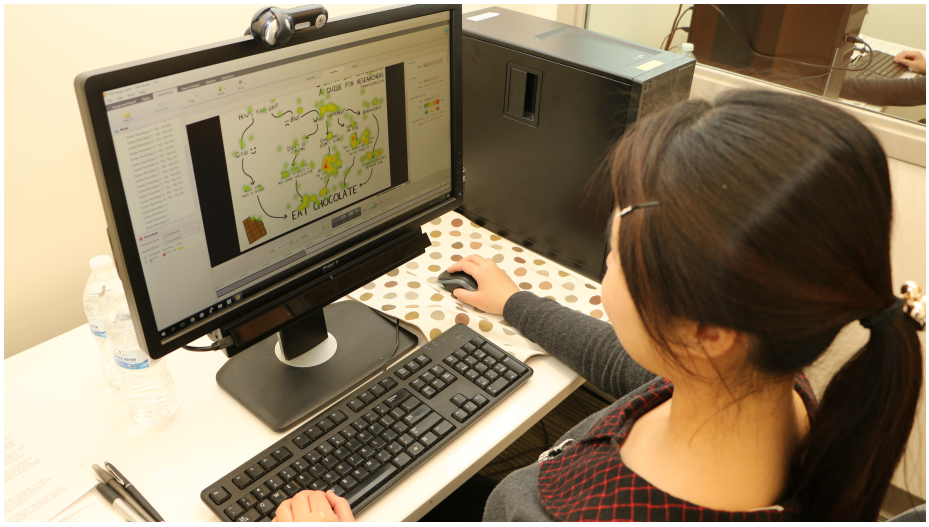


Figure 3.7. Experiment setup of the follow-up study. Each participant conducted think-aloud sessions in front of the monitor of a desktop PC. A web camera (audio recording only) is placed on top of the monitor, and an eye-tracker is attached at the lower part. Photo taken by DeEtte Starr.

resource during coding. Interview responses were considered when generating the models of AE. The data analysis was performed in two main stages:

1. A qualitative coding stage, in which three qualitative coding methods were employed; and

2. A triangulation stage, in which data from mixed methods was analyzed and cross-referenced with the findings from the coding stage.

In the qualitative coding stage, we selected verbal protocol data from the think-aloud sessions as our primary target of analysis. The researchers used a verb + ing style to describe the behaviors or cognitive activities in which participants were engaged during in the transcription. Names of the activities were mostly adapted from the codes, but some of them were modied.

Three qualitative coding method were explicitly used to identify patterns and themes within the data: **Process coding**, **Emotion coding** and **Causation coding** [152]. We used process coding to identify the events or activities carried out

through the session; Emotion coding was used for identifying users' emotions and their affective reactions during the session. Finally, the Causation coding was used to identify the "chains" of codes from the previous two procedure and identify causal relationship between them.

The same three coders continuously had discussions during the coding process to ensure agreement. If there were disagreements, the primary researcher (me) helped resolve it. Several sessions were conducted to triangulate the coders styles and strategies. During each coding phase, the researchers created affinity diagrams to help categorize the codes and identify similarities and differences among them. Understanding of the development and structure of affective engagement in InfoVis was gradually generated throughout the process, and nalized at the end of the selective coding stage.

3.2.3 Qualitative Coding

From this iterative coding process, the coders came up with a code book which comprises two primary categories: emotion and process. In the emotion codes been identified from emotional coding, there are 4 subcategories, under each of them, there are codes belong to it:

- Happy (Enjoyment, Excitement, Expressing likeness, Fun, and Relaxing),
- Unhappy (Annoyed, Disappointed, Frustrated, Loss of Trust, Not satisfying, and Uninterested),
- Surprise (Mild surprise, No surprise, and Positive Surprise),
- Troubled (Confusion, Concerning, Helpless, and Wondering); and
- Two other codes that didn't fall into these categories (Careless and Curious).

Following similar approach, from process coding, we identified 6 sub-categories and codes under each of them:

- Asking (Assuming and Questioning),
- Wondering (Being confused, and Hesitating),

- Explore (Investigating, Searching, and Testing),
- Knowledge (Learning new things,and Recalling),
- Making sense (Interpreting, Sensemaking, and Summarizing),
- Collecting data (Reading and Skimming); and
- Four other codes that didn't fall into these categories (Creating, Focusing, Judging, and Rationalizing)

3.2.4 11 Indicators of Affective Engagement

Following a similar triangulation process to the previous experiment, this time the focus was on characterizing the behaviors and activities that are associated with emotional aspect of engagement. At the end, 11 indicators of AE were proposed, for each of them coders wrote down the description and its reference codes from our qualitative analysis results, and then assigned a label that summarize its main idea:

- **Fluidity**–The flow of use is continuous, smooth, not disrupted. This indicator was derived from the chains of participants' activities where an interruption of "flow" happened, such as *being confused* or *hesitating* while working with a visualization, resulting in unhappy reactions such (e.g., *disappointed* or *annoyed*. These kinds of interrupting experiences should be avoided.
- **Enthusiasm**–The user is interested, eager, proactively involved, and motivated. This indicator is strongly related to exploratory types of actions such as proactively *searching* and *testing*, resulting in happy-ish emotions such as *excitement* and *fun*; It also sometimes associated with surprise-related codes such as *surprise (positive)*.
- **Curiosity**–The user is inquisitive or investigative. This indicator is derived from codes such as *curious* and activities under the exploration category such as *Investigating*.
- **Discovery**–The user discovers something new or noticeable. This indicator is derived from instances where users expressed gaining new knowledge, making

connections with existing knowledge, or wanting to know more about the content of a visualization. Those activities often time were coded into process such as *Learning new things* and *Recalling*, followed by emotions such as *curious* and *surprise*.

- **Clarity**–The visualization express a clear concept or message; the story is clear or can be easily interpreted. This indicator is derived from instances where users could understand what the visualization was intended to convey, they shows positive emotion such as codes under *happy* subcategory after activities been coded as *Interpreting* or *Summarizing*.
- **Storytelling**–The visualization tells a compelling story or has a persuasive narrative style. This indicator is derived from codes such as *sensemaking* and *summarizing*. Participants usually demonstrated a more absorbed status such as *focusing* when they are into a story wither from the content or from their own memory.
- **Creativity**–The visualization helps users generate or express creative and innovative thoughts or ideas. This indicator is derived from instances where users were imagining new ideas as a result of seeing a visualization. And activities codes as *creating* where they activity Coming up with some ideas, followed by positive emotions (e.g., codes under *happy* category) that will further encourage they to create more.
- **Entertainment**–The user feels the visualization is fun, interesting, or charming. This indicator is derived from instances where users expressed feeling happy, emotions such as *Enjoying* and *Exciting* can be easily identified from both think-aloud and the tone of speaking.
- **Untroubling**–The user feels content and does not feel upset or frustrated. This indicator is derived as a reverse code from instances where users expressed feeling *confused*, *annoyed*, or *helpless*.
- **Captivation**–The user is concentrating, absorbed. This indicator is derived from instances where users felt like they *lost track of time* or *forgot about their*

surroundings while using visualizations, and often time been coded as *focusing* during their explorations (e.g., *Investigating*, *Searching*, and *Testing*).

- **Pleasing**–The user is impressed by the visualization (e.g., visually, concept-wise). This indicator is derived from instances when users expressed feeling pleased and happy with the design of the visualization.

Three Levels of Affective Engagement Indicators

From the different characteristics we observed from the participants, naturally, we categorized the 11 elements into three levels from the patterns emerged during qualitative coding process:

- **Process** (Fluidity, Enthusiasm, Curiosity, and Discovery): where people present a specific patterns of behaviours. In causation coding, it usually shown as a chain of activities, sometimes we can even see loops among those chains.
- **Judgment** (Clarity, Storytelling, and Creativity): where people provide their personal decisions or opinions based on their logical reasoning. It can be told through process coding by identifying user's judgmental phases such as "I would say....is...."
- **Feeling** (Untroubling, Captivation, and Pleasing): where people show their emotional reaction. In emotional coding, either participants will specify their current emotion, or it can be identified from participants tone or the pattern of their mouse moving.

3.3 Summary

In summary, to elicit and characterize Affective Engagement(AE) in InfoVis, two mixed-method lab studies were conducted where we asked our participants to interact with various visualizations, and captured participants' affective engagement. From an iterative coding process, two AE models as our theoretical background of AE were proposed. One is the structure and characteristic model, and other one is the

process and development model. In the follow-up study, a similar experiment was conducted but with different tested visualizations and coding strategy. In the end, 11 indicators of AE are proposed through this process: *Fluidity, Enthusiasm, Curiosity, Discovery, Clarity, Storytelling, Creativity, Entertainment, Untroubling, Captivation, and Pleasing*, in which they are categorised into three levels: behavior, judgment, and feeling (see Table 3.1).

An interesting observation that emerged from the 11 indicators is that, some of the indicators are like person traits (largely stable within persons, e.g., enthusiasm, curiosity, creativity), while others as properties of the interaction between person and the visualization (probably less stable within a particular person, e.g., clarity, storytelling). This might be an alternative way to consider the underlying construct of AE; the internal person trait such as people's personalities do influence their subjective feelings and judgments One the other hand, other indicators are like external factors or variables only existed during the interaction–they are more "situated" and contingent (i.e. [153]).

The primary purpose of conducting two mixed method investigations is to establish the theoretical background of AE. As elaborated in the literature review section, the term "engagement" is widely used in our daily life, hence the operational definition of AE needs to be characterized through this grounded approach. The results in this Stage 1 can contribute to the development of AE assessment instrument in the next stage as the content of our target construct (AE) and the references of AE's behavior domain. Moreover, the alternative AE construct mentioned above can be examined using analytical methods in the field test, which would be described in detail in Stage 3.

Table 3.1.

11 indicators of AE categorized into three levels.

Category	**Indicator**	**Description**
Behavior	Fluidity	The flow of use is continuous, smooth, not disrupted
	Enthusiasm	The user is interested, eager, proactively involved, and motivated
	Curiosity	The user is inquisitive or investigative
	Discovery	The user discovers something new or noticeable
Judgment	Clarity	The visualization express a clear concept or message; the story is clear or can be easily interpreted
	Storytelling	The visualization tells a compelling story or has a persuasive narrative style
	Creativity	The visualization helps users generate or express creative and innovative thoughts or ideas
Feeling	Entertainment	The user feels the visualization is fun, interesting, or charming
	Untroubling	The user feels content and doesn't feel upset or frustrated
	Captivation	The user is concentrating, absorbed
	Pleasing	The user is impressed by the visualization (e.g., visually, concept-wise)

4. STAGE 2–DEVELOP ASSESSMENT OF AFFECTIVE ENGAGEMENT

As concluded at the end of the introduction section, an evaluation tool of AE in InfoVis domain need to be developed. After the theoretical background and characteristics of AE been established in stage 1, in this section, we will develop the self-report survey instrument that can assess a users level of AE.

Since AE is a complex and unobservable (latent) construct, in order to be able to measure or assess it, there are several general procedure that can be utilized in the construction of such assessment instrument [31, 154–156]:

- First, the (latent) target construct needs to be decomposed into several subcomponents/key factors based on its conceptual space.
- Second, the observable behavior indicators need to be established for each subcomponent or key factor.
- Finally, based on those behavior indicators, items (questions in the survey) can be written that can locate respondents' level on the indicator scale.

A structural overview of the survey instrument for affective engagement in InfoVis is shown in Figure 4.1, from top to bottom: Target construct (affective engagement), Key factors (denoted as K), Behavior indicators (denoted as B), and Items (denoted as X). As it is possible to have multiple key factors for a construct, it is also possible to have multiple behavior indicators and items from their previous layer.

There is a general procedure to develop an instrument to assess an affective construct [31, 156–158]. Three primary steps of measurement development need to be employed:

1. Identify intended use of test scores;
2. Interpret scores related to AE and test-taker population(s); and

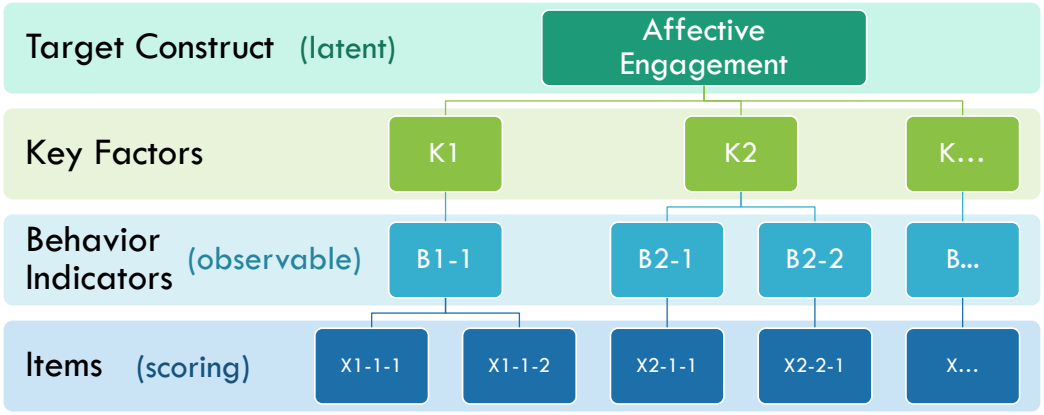


Figure 4.1. Structural overview of the survey instrument for affective engagement in InfoVis.

3. Develop a conceptual definition (space) of the target construct and establish a list of behaviors that are taken to indicate a person's location on the trait continuum.

4.1 Instrument Specifications

There are 11 selected-response items in total using Likert scales [159]. Each item is a statement that needs the respondent to answer with a degree of agreement (from strongly agree to strongly disagree). When using Likert scales where all options have labels, 5 and 7 response options are generally suggested from the literature. However, as Bandalos [160] points out, the ability of respondents to reliably distinguish the scale points depends on their level of education; thus, in the general population, where respondents' educational levels are unknown in advance, using fewer scale points is a reasonable choice. Therefore, we adopted five-point Likert scale for our instrument.

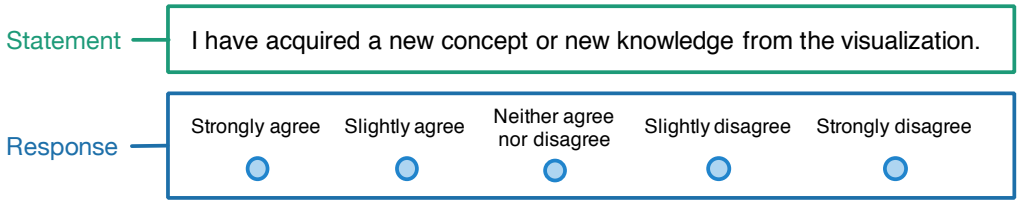


Figure 4.2. An example item of AEVis, there are two components for each of the item: Statement and Response.

Table 4.1.
Scoring system of AEVis which adopted five-point Likert scale.

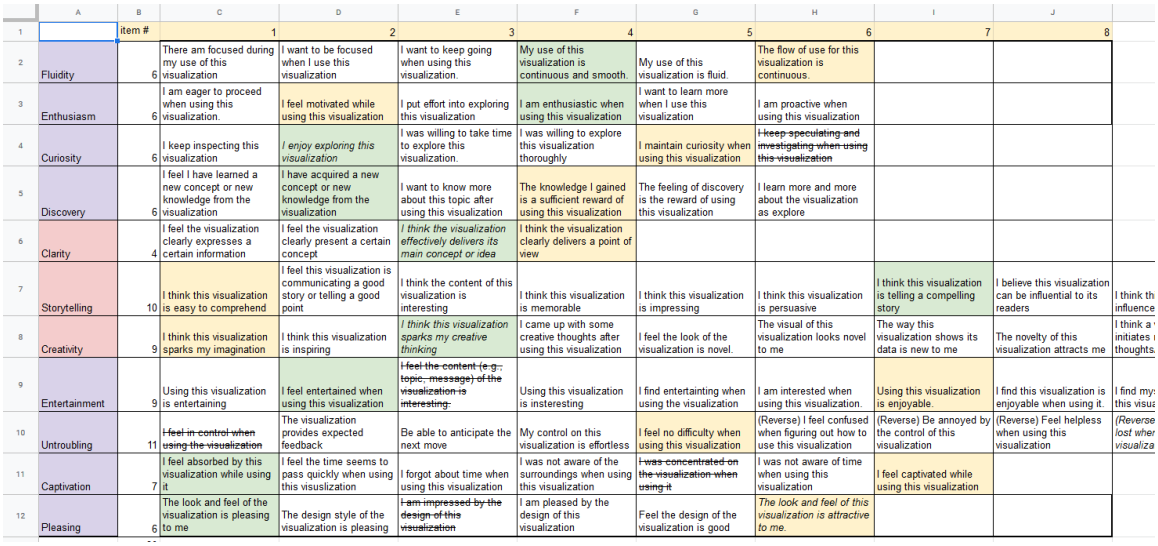
5 points	**4 points**	**3 points**	**2 point**	**1 point**
Strongly agree	Slightly agree	Neutral	Slightly disagree	Strong disagree

The total score for each respondent is computed by adding all points together. As shown in Table 4.1, each selected-response item is given 1-5 points (5 points for strongly agree, 4 points for slightly agree, 3 points for neither agree or disagree, 2 points for slightly disagree, and 1 point for strongly disagree). The maximum score is 5 (points) per item—thus, the maximum total score is 55 for all 11 items while the minimum score is 11.

4.2 Item Writing

With the instrument specifications in mind, the item writing was began, which is a process of transferring sub-components into questions (i.e., "items"). Based on the 11 AE indicators and 3 levels theory established from the previous stage, the research team had multiple brainstorming sessions and then I drafted a total of 137 potential items (see Figure 4.3).

The complete item bank (or refer as "item pool") can be found in the Appendix. For indicators under **behavior** level, candidate items tend to use active verbs to



		item #	1	2	3	4	5	6	7	8	
Fluidity	6	There am focused during my use of this visualization	I want to be focused when I use this visualization	I want to keep going when using this visualization.	My use of this visualization is continuous and smooth.	My use of this visualization is fluid.	The flow of use for this visualization is continuous.				
Enthusiasm	6	I am eager to proceed when using this visualization.	I feel motivated while using this visualization	I put effort into exploring this visualization	I am enthusiastic when using this visualization	I want to learn more when I use this visualization	I am proactive when using this visualization				
Curiosity	6	I keep inspecting this visualization	*I enjoy exploring this visualization*	I was willing to take time to explore this visualization.	I was willing to explore this visualization thoroughly	I maintain curiosity when using this visualization	~~I keep speculating and investigating when using this visualization~~				
Discovery	6	I feel I have learned a new concept or new knowledge from the visualization	I have acquired a new concept or new knowledge from the visualization	I want to know more about this topic after using this visualization	The knowledge I gained is a sufficient reward of using this visualization	The feeling of discovery is the reward of using this visualization	I learn more and more about the visualization as explore				
Clarity	4	I feel the visualization clearly expresses a certain information	I feel the visualization clearly present a certain concept	*I think the visualization effectively delivers its main concept or idea*	I think the visualization clearly delivers a point of view						
Storytelling	10	I think this visualization is easy to comprehend	I feel this visualization is communicating a good story or telling a good point	I think the content of this visualization is interesting	I think this visualization is memorable	I think this visualization is impressing	I think this visualization is persuasive	I think this visualization is telling a compelling story	I believe this visualization can be influential to its readers	I think th influence	
Creativity	9	I think this visualization sparks my imagination	I think this visualization is inspiring	*I think this visualization sparks my creative thinking*	I came up with some creative thoughts after using this visualization	I feel the look of the visualization is novel.	The visual of this visualization looks novel to me	The way this visualization shows its data is new to me	The novelty of this visualization attracts me	I think a initiates thoughts	
Entertainment	9	Using this visualization is entertaining	I feel entertained when using this visualization	~~I feel the content (e.g., topic, message) of the visualization is interesting.~~	Using this visualization is insteresting	I find entertainting when using the visualization	I am interested when using this visualization.	Using this visualization is enjoyable.	I find this visualization is enjoyable when using it.	I find my this visu	
Untroubling	11	~~I feel in control when using the visualization~~	The visualization provides expected feedback	Be able to anticipate the next move	My control on this visualization is effortless	I feel no difficulty when using this visualization	(Reverse) I feel confused when figuring out how to use this visualization	(Reverse) Be annoyed by the control of this visualization	(Reverse) Feel helpless when using this visualization	*(Reverse lost wher visualiza*	
Captivation	7	I feel absorbed by this visualization while using it	I feel the time seems to pass quickly when using this visuslization	I forgot about time when using this visualization	I was not aware of the surroundings when using this visualization	~~I was concentrated on the visualization when using it~~	I was not aware of time when using this visualization	I feel captivated while using this visualization			
Pleasing	6	The look and feel of the visualization is pleasing to me	The design style of the visualization is pleasing	~~I am impressed by the design of this visualization~~	I am pleased by the design of this visualization	Feel the design of the visualization is good	*The look and feel of this visualization is attractive to me.*				

Figure 4.3. Screenshot of the item bank for this study. Each row contains all item candidates for one of the 11 indicator. Cells colored in yellow and green are items been selected for the expert review process.

describe user's actions such as "I want to ", "I put effort", and "My use of this visualization is". For indicators under **judgement** level, items tend to use "I think" and "I believe" to initiate the description of a status or quality of an experience around using a visualization. Finally, for indicators under **feeling** level, items tend to initiate the description of an emotion status or a sentiment aspect from phrases such as "I feel" and "I am aware".

As the main item writing contributor, I considered two main criteria when writing our items—precision and conciseness:

- By **precision**, it means the language should point to the exact meaning while avoiding ambiguity. The statement should not have grammatical conjunctions such as asking multiple things at the same time, and avoid inappropriate adverbs (e.g., modifiers, intensifiers) or verbs that are defined vaguely by nature.
- By **conciseness** , it means the item should be brief but comprehensive. A respondents cognitive load can be unnecessarily increased by asking complex or verbose sentences.

Plus, when coming up with the corresponding behaviors for the 11 AE indicators, I decided not to include specific physical components of visualizations (e.g., visual encoding, spatial organization, visual representation) in the statements. Instead, the writing was oriented around user's subjective experience such as how they act, react, and feel about the visualizations. Due to the fact that not all visualizations contain the same elements or components, some items might not be able to adapt various types of visualizations if they contain exact descriptions of those elements.

Some variations in the items were established at this stage. Since the term "visualization" can be interpreted in different ways, we leave visualization as a placeholder and suggest the survey administrator to insert the most contextually appropriate term–e.g., diagram, chart, graphic, and so on. As for static visualizations that are passively viewed, the terms "use" and "using" in some items did not seem appropriate; thus, "use" and "using" can be swapped with more appropriate terms such as "view" and "viewing".

4.3 Tryout Sessions

To ensure a survey instrument appropriately represents the construct space, several tryout sessions with target respondents of the instrument are recommended. I recruited a small group of participants to help me identified issues in the early stage of development. The respondent population is general public with adequate level of English literacy as the intended use scenario of AEVis is to measure user's affective engagement with communicative or narrative type of visualizations, specifically for non-expert topics and tasks. There is no particular criteria or restriction on the participants in the tryout session except being able to read and interpret the items written in English.

The primary purpose of tryout session is to to solicit respondent's possible cognitive process when they went through the items in choosing their response from among the options. It is a rather exploratory process, therefore formal usability eval-

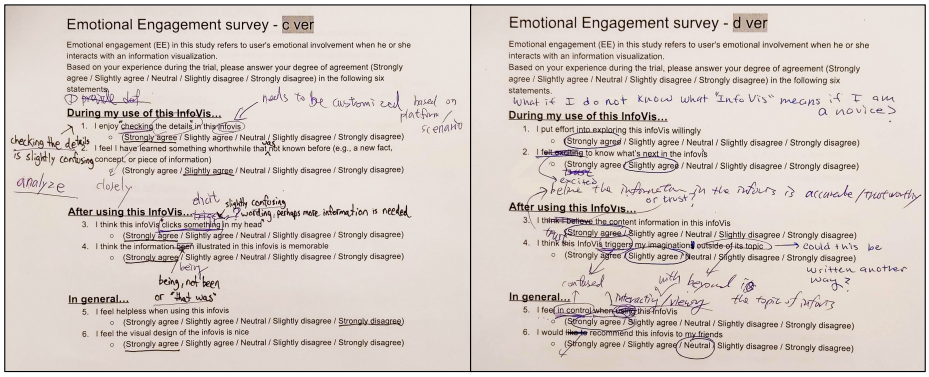


Figure 4.4. Two examples of instructions and six items for tryout sessions. Both participants and a researcher made notes or suggestions on the print out.

uation methods such as cognitive walkthrough [161] or verbal protocol [162] were not suitable for this purpose. During the tryout sessions, participants are encouraged to freely speak out their thoughts and share the way they interpreted the provided item statements. Furthermore, I would stay with the participants to answer their questions or even to have conversations with them.

User review from tryout sessions were collected from five Purdue graduate students (two in Technology, two in Computer Science, one in Politics; two of them are female) and two Purdue undergraduate students (one in Technology, one in Engineering; one of them is female). Participants were contacted and scheduled with me individually. In each tryout session, a print out or a .pdf file contains the instrument instruction and the selected items from the item bank was given to the participant, numbers of items are varied depending on the participants' available time. The participants have to use a selected interactive visualization that we collected in stage 1, and then provide their responses on the selected AEVis items. Two examples of instructions and items for tryout sessions are shown in Figure 4.4.

Based on participants' suggestions and comments, the research team reduced the number of item candidates from 137 to 80 items, the updated item bank can be found in the Appendix (see Table B.6, B.7, and B.8). Beside minor modifications such as

wording across all indicators, several new items have been added into Fluidity and Storytelling two indicators. On the other hand, numbers of items have been removed from the rest of nine indicators. The change of item number across 11 indicators is summarized in Table 4.2.

Table 4.2.
The change of item numbers across 11 indicators before and after the try-out sessions, the difference for each indicator is calculated by subtracting number of items before and after the tryout sessions.

		Number of Items		
Category	**Indicator**	**Before tryout**	**After tryout**	**Difference**
Behavior	Fluidity	5	6	+1
	Enthusiasm	9	6	-3
	Curiosity	15	6	-9
	Discovery	19	6	-13
Judgement	Clarity	16	4	-12
	Storytelling	5	10	+5
	Creativity	18	9	-9
Feeling	Entertainment	16	9	-7
	Untroubling	13	11	-2
	Captivation	10	7	-3
	Pleasing	11	6	-5

4.4 Expert Review

Subsequent to the user review in the tryout sessions, an expert review was conducted with the aim of achieving a reasonable level of content validity [156,163]. After several internal meetings within the research team, 22 items were selected for expert review, where each indicator had two candidate items (see Table 4.3).

Table 4.3.
The 22 candidate items and the item assignment that were used in expert review process. Each of the 11 indicators have 2 corresponding items. The ✓ in the cells indicate the domain expert who reviewed the corresponding items.

		Expert		
Indicator	**Item Statement**	**no.1**	**no.2**	**no.3**
Fluidity	My use of this visualization is continuous and smooth.			✓
	The flow of use for this visualization is continuous.	✓	✓	
Enthusiasm	I am enthusiastic when using this visualization			✓
	I feel motivated while using this visualization	✓	✓	
Curiosity	I enjoy exploring this visualization			✓
	I maintain curiosity when using this visualization	✓	✓	
Discovery	I have acquired a new concept or new knowledge from the visualization	✓		✓
	The knowledge I gained is a sufficient reward of using this visualization		✓	
Clarity	I think the visualization effectively delivers its main concept or idea	✓		✓
	I think the visualization clearly delivers a point of view		✓	
Storytelling	I think this visualization is telling a compelling story	✓	✓	
	I think this visualization is easy to comprehend			✓
Creativity	I think this visualization sparks my creative thinking	✓	✓	
	I think this visualization sparks my imagination			✓
Entertaining	I feel entertained when using this visualization	✓	✓	
	Using this visualization is enjoyable.			✓
Untroubled	(reverse coded) I feel frustrated when using this visualization		✓	
	I feel no difficulty when using this visualization	✓		✓
Captivation	I feel absorbed by this visualization while using it		✓	
	I feel captivated while using this visualization	✓		✓
Pleasing	The look and feel of the visualization is pleasing to me		✓	
	The look and feel of this visualization is attractive to me.	✓		✓

The study was conducted through Qualtrics online survey platform [164] (IRB protocol#: 1902021654). Participants were recruited through a recruitment e-mail which contains a link to our online survey system. After agreeing to the consent form, participants were then directed to an instruction page that contains the purpose and process of the study, an external website with a brief overview and the descriptions

of current 11 indicators was provided as well. In the next page, participants were asked to fill out a survey consists of demographic information, their opinions toward provided survey items, and comments for each of the items. After completing the survey, participants were compensated by $10 giftcard, is paid through e-mail.

Indicator: Enthusiasm - The user is interested, eager, proactively involved, and motivated Statement: "I am enthusiastic when using this visualization."

- **Cla rity:** How clear is this statement (e.g., conciseness, readability, understandability)?
  - Completely Unclear
  - Somewhat Unclear
  - Very Clear
- **Relevance:** How relevant is this statement to the indicator "Enthusiasm"?
  - Completely Unclear
  - Somewhat Unclear
  - Very Clear
- **Comments:** If you selected "unclear" or "irrelevant" in the above questions, please elaborate to help us understand why you think so, and what improvements could be made. If you selected "relevant" or "clear", please elaborate on what specifically makes the statement good. Other suggestions are also welcome (e.g., suggest changes in wording or suggest that the item be eliminated).

Figure 4.5. An example question set (indicator: enthusiasm) in an expert review survey form.

The invited experts were person who had considerable experience in InfoVis and were active in teaching and research. As Lynn suggested, a small group of domain experts (e.g., three) can be sufficient [163]. Domain experts were invited via Dr. Paul

Parsons' personal network and three of them responded to the expert review forms. All recruited domain experts are InfoVis researchers who worked in research universities (schools with high research activity). The first domain expert is an associate professor with 18 years of experience in InfoVis, his areas of research are focused on InfoVis, Visual Analytics, and HCI. The second domain expert is an associate professor with 8 years experience, his research areas are HCI, Immersive Systems, and InfoVis. The third domain expert is an instructor with more than 6 years of experience in InfoVis, her research areas are InfoVis and CS (Computer Science) education.

Generally, several structural elements are suggested to be included in a domain expert review [156, 165]: Item content (whether the content domain adequately measures all dimensions of the construct), Item style (whether the item construction and wording are clear), and Comprehensiveness (whether the all items in the instrument properly cover the content domain). Therefore, during this experiment, experts were asked to review and provide their judgments on items **clarity**, **relevance**, and a **short comments**.

As shown in Figure 4.5, for each indicator, the domain expert provided two multiple choice responses and one short response questions. Item's clarity means people's perception of the clarity of the item (conciseness, readability, understandability); item's relevance means how relevant a person feels each item is to the construct. Based on experts' responses of the two multiple choice questions, they were encouraged to specify the reasons and to suggest improvement or revisions of the item. An example of the survey instruction and two demographic questions can be found in Appendix (See Figure B.3, B.5, and B.6).

Each domain expert received an expert review survey form with different items in it. To make sure that domain experts could have a better understanding on the whole picture of AE, when designing the expert review survey forms, I intentionally assigned at least one item from each indicator so that every domain expert could assess all 11 AE indicators. Also, each of the items could be reviewed by at least one of the domain expert. The item assignment are indicated in Table 4.3.

Table 4.4.
The summary table of expert feedback, text in each cell indicates the type of suggestions domain experts provided to the indicators.

Indicator	**Expert no.1**	**Expert no.2**	**Expert no.3**
Fluidity	clarity		
Enthusiasm			clarity
Curiosity	relevance		
Discovery			
Clarity	clarity	clarity	
Storytelling			
Creativity		wording	clarity, relevance
Entertaining			
Untroubled		clarity	relevance
Captivation	relevance		
Pleasing	relevance		wording

At the end, the three domain experts provided different types of suggestions on the 11 indicators; some of them were more related to the clarity of the items, some of them were more concerned about the relevance between the indicator and the definition of AE. A summary table regarding the types of comments they given can be found in Table 4.4. With help from three domain experts, several items were modified (e.g., "I feel immersed while using this visualization" might make people think about virtual reality, so the "immersed" was changed to "absorbed") and eliminated some of the inappropriate items.

4.5 Findings

Based on the theoretical background of AE and 11 indicators established in stage 1, an item bank contains 137 item candidates was developed. Then, with help of

7 target population users, we revised and decreased the size of item bank into 80 items and selected 22 items that could be reviewed by InfoVis domain experts. After expert review with three domain experts, the most appropriate item from 11 pairs of candidates were picked. By grouping items based on the three categories (behavior, judgment, and feeling), the first version of the survey instrument was assembled. The developed instrument had 11 items, each corresponding to one indicator (See Table 4.5).

Table 4.5.
The first version of AEVis, there are 11 items in it, the order of the items is based on the 3 levels theory developed in stage 1.

#	**Item Statement**
Q1	My use of this visualization is continuous and smooth.
Q2	I feel motivated while using this visualization
Q3	I enjoy exploring this visualization
Q4	I have acquired a new concept or new knowledge from the visualization
Q5	I think the visualization effectively delivers its main concept or idea
Q6	I think this visualization is telling a compelling story
Q7	I think this visualization sparks my creative thinking
Q8	I feel entertained when using this visualization
Q9	I feel frustrated when using this visualization (reversed coded)
Q10	I feel absorbed by this visualization while using it
Q11	The look and feel of the visualization is pleasing to me

5. STAGE 3–FIELD TEST OF AEVIS

Once the first version of AEVis survey instrument been developed, the next step is to test and revise the instrument in a field test. The primary purpose of this step in instrument development is to collect responses for further analysis that can contribute to instrument revisions and score interpretation validations. The study was conducted on Amazon Mechanical Turk, three rounds of field tests were established to serve our purpose for this stage.

5.1 Experiment

The field test in stage 3 was done by testing users level of engagement with provided visualizations, and observe how current AEVis and its items performed (IRB #: 1611018468). Since responses from the intended target population (i.e., the general public) is preferred, participants were recruited through Amazon Mechanical Turk [166] (MTurk) where has easier access to the general public population [167]. When participants finish a HIT (Human Intelligence Task) in MTurk, they could be approved by the researchers and then be automatically get paid through the platform. The participants will be compensated with $2.00 for the experiment. If they choose to quit the experiment before completion, they will not get their compensation. By assigning user qualification in the MTurk control panel, the participants will only be recruited once. Therefore, the research team can prevent participants from submitting multiple responses.

The experiment was hosted to Qualtrics online survey platform [164]. Participants would first read an announcement explaining the goal and concept of the study. After agreeing to participate, participants were redirected to the online survey system, where they were presented with an online consent form and a demographic survey.

Next, each participant have to complete 3 trials, where each trial included a tested visualization, a set of content-related questions, and the AEVis survey containing 11-13 items (see Table 5.1). The order of the trials was randomized.

To increase the likelihood that participants actually explored the visualizations (rather than clicking randomly just to get paid), we designed three content-related questions for participants to answer before taking the AEVis survey as suggested by Thomas [168]. For each trial, the first question was about the topic of the visualization (e.g., what this interactive chart is about?); the second question is about eliciting or reading a specific value or piece of information from the visualization (e.g., Technology that numbered as 35 is?); the third one is about retrieving a specific value/number from the visualization (e.g., How many goals did Italy score in 2002's tournament?). The first two were multiple-choice questions, while the third was a short response question. All experiment materials such as recruitment announcement and screenshot of online survey can be found in the appendix.

Besides implementing three content-related questions for each visualization to ensure participants actually explore the tested visualizations, the recorded time spent is also utilized to filter out people who not paying attention to the tasks. Since items in the survey instrument require respondents to read a statement and rate their degree of agreement on a Likert scale, based on the on-site pilot tryout, we believe a reasonable estimated time spent for completing the survey instrument once should take at least approximately 15-20 seconds. Therefore, responses that spent less than 10 seconds for more than two out of the three trials were dropped.

5.1.1 Tested Visualizations

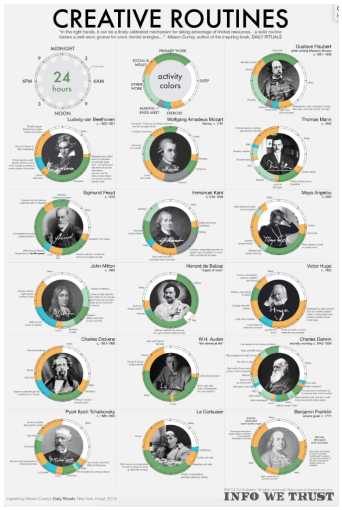
By testing AEVis on three visualizations in field tests, our aim was to investigate how the developed instrument and items behaved with different data, encodings, layouts, and so on. The three chosen visualizations were not intended to be representative of all visualizations. Plus, it is practically not possible to exhaustively cover all

Table 5.1.
13 items been tested in the field tests. The indicator column specify items corresponding AE indicators while the category is the characteristics of that AE indicator. The order (#) is the item sequence in the final version of AEVis. Note that Untroubling v1 and v2 didnt make into the final instrument and therefore have NA in the order columns.

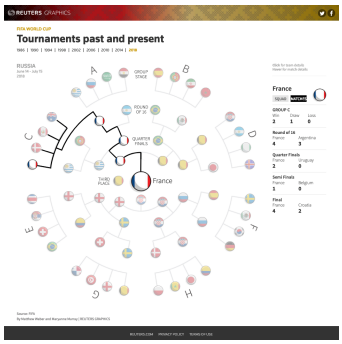
Category	**Indicator**	**#**	**Item Statement**
Behavior	Fluidity	1	My use of this visualization is continuous and smooth.
	Enthusiasm	2	I feel motivated while using this visualization
	Curiosity	4	I enjoy exploring this visualization
	Discovery	5	I have acquired a new concept or new knowledge from the visualization
Judgment	Clarity	6	I think the visualization effectively delivers its main concept or idea
	Storytelling	7	I think this visualization is telling a compelling story
	Creativity	8	I think this visualization sparks my creative thinking
Feeling	Entertainment	9	I feel entertained when using this visualization
	Untroubling v1	NA	I feel **frustrated** when viewing this visualization.
	Untroubling v2	NA	I feel no difficulty when viewing this visualization.
	Untroubling v3	3	I **don't** feel frustrated when using this visualization
	Captivation	10	I feel absorbed by this visualization while using it
	Pleasing	11	The look and feel of the visualization is pleasing to me

types of visualizations in the field test. Hence, two dimensions were identified to help selecting the visualizations that are reasonably different from each other. First, since size/scale is one of the primary characteristic of visualizations, orientated by our experiment platform desktop PC, the research team defined three levels of visualization size: partial screen, full screen, more than 1 screen. Second, since the interactive features would significantly influence how an user perceive and use a visualization, the research team defined three interaction levels for visualizations: static (no interaction), basic (simple click and hover), and advanced (filtering, zooming, brushing, etc.).

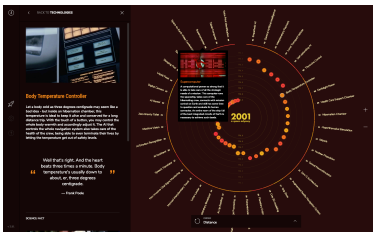
After reviewing visualizations that are freely available on the web, 23 candidate visualizations were collected and their level of size and interaction levels were identified one by one (See appendix for the full list). At the end, three visualizations were



(a) "Creative Routines" [169]



(b) "FIFA World Cup" [170]



(c) "Space Odyssey" [171]

Figure 5.1. The three initial selected tested visualizations for field test. Note that the tested visualization no. 2 has been changed in the second round of field test.

chosen where each one satisfied one of the level in each dimension. The first tested visualization [169] has more than 1 screen size and static interaction level; The second tested visualization [170] has partial screen size and basic interaction level; the third one [171] has full screen size and advanced interaction level (See figure 5.1). While this is far from an exhaustive sample of different visualization types and features, we believe it offers a reasonable degree of variation for testing our instrument.

5.1.2 Analytical Methods

Three analytical methods were utilized for analyzing the collected data: item analysis, factor analysis, and IRT (Item Response Theory). Specifically, item analysis has been performed in the 1st and 2nd round of the field test; EFA (Exploratory Factor Analysis), CFA (Confirmatory Factor Analysis), and IRT were performed with the dataset combining the second and the third round of field test.

Item Analysis is the first analytical method to be used in the field test, it is a set of statistical procedures to identify problematic or biased test items, the patterns and characteristics of the item responses were examined during this process [172]. The practical reason to conduct item analysis first is that it can reveal potential issues of AEVis items as early as possible and be performed with a relatively smaller number of responses. Statistical procedures to identify invalid biased test items are often sophisticated and costly to test developers. For example, Nunnally suggested to have an initial pool be composed of 1.5 to 2 times as many items as the final instrument [173]. Thus, an alternative approach is to employ instrument review by domain experts and identify potential invalid or biased items beforehand [174]. Some debates on the balance between the two approaches were studied as well [175].

Therefore, an alternative strategy were presented for the generation of items. Nunnally and Bernstein suggested to construct a relatively smaller number of items first (e.g., 30 when 40 is required to obtain coefficient Alpha of 0.8) [176]. These items then would be pilot tested using a smaller sample size, if an adequate level of reliability coefficient can't be obtained, additional items would be constructed, and the updated items would have another round of pilot-test with another group of participants. The researcher would iterate this process until an adequate level of reliability is achieved. This strategy is labor intensive, but the early results can reveal potential issues, then the researchers would be able to stop or revise at any time, thus saving time, effort, and budget. In this field test, due to time and budget constraints, a similar strategy was adopted where the pilot testings with smaller sample size would be continued until all noticeable issues have been resolved via item analysis.

Factor Analysis including EFA and CFA were conducted after item analysis. Factor analysis have been widely used in theory development and assessing construct validity [177, 178]. This latent variable models is often used to explain a larger set of j measured variables with a smaller set of k latent constructs; in instrument development, it means the constructs or factors can be used to represent the observed

scores [179]. Typically, these techniques require larger number of responses which suggested more unique respondents to be recruited, and usually cost significantly more money when collecting data (discussions can be found in [180–183]). Thus, I decided to perform these two analytical methods in the 3rd round of field tests with more responses. Normally, people would conduct EFA to explore possibilities of latent variables in the model, and to determine the number of factors to extract. After reviewing alternatives under different model parameters and proposing possible explanations of the models. Then, CFA can be performed to determine whether the proposed models and their interpretations were appropriate or not, the model fit can be calculated with the number of factors being determined on the basis of previous exploratory studies [184,185]. It is believed that following an EFA with a CFA on the same data set can be potentially misleading [179]. Due to the resource constraints such as budget and time limit, the research team examined an existing dataset in order to evaluate the construct validity of this instrument using CFA.

Item Response Theory was the last analytical method to be conducted in the filed test, to reexamine the construct validity of the AEVis after the results from factor analysis has been established. Generally, IRT models requires more responses than classical test theory (CTT) statistics, previous studies have provided various suggestions based on both simulation and empirical studies (e.g., [186–188]). IRT can be used to evaluate the psychometric properties of an existing scale and its items, and generates rich item level information and offers many advantages over CTT [189,190]. Plus, the assumptions about the scale of measurement for each response distribution in linear FA (e.g., EFA, CFA) are different from the nonlinear one (e.g., IRT); because the empirical response distributions for AEVis items were notably skewed in the field test (participants tend to report higher scores), the IRT assumptions seems to be more appropriate than the FA ones Therefore, IRT method was used in a complementary manner of the previous factor analysis.

5.2 Field Test Round 1

As the first part of field test, 50 participants were recruited from MTurk. After removing responses from 2 participants based on the criteria mentioned above, two of them were dropped and the dataset ended up with 48 participants. Each participant took 3 trials, thus there were a total of 144 survey responses. The average of responses collected in field test round 1 are shown in Table 5.2, which including average total score, total time spent on trial, and total time spent on the instrument for the three tested visualizations.

Table 5.2.
Average of total score, total time spent on trial, and total time spent on the instrument for the three tested visualisation in field test round 1, include all 3 trials, N = 48.

	Viz no1	**Viz no2**	**Viz no3**
Total Score (point)	42.02	41.49	41.6
Time Spent on Trial (sec)	134.15	101.02	239.73
Time Spent on Instrument (sec)	33.92	30.90	34.04

To identify whether survey items in the first version of AEVis performed as expected, an traditional CTT item analysis was conducted using R packages Psych [191] and PolycorR [192]. There are some general guidelines can be adopted in the domain of instrument development (see [154, 157]). For each item, I calculated and examined its:

1. Classical item difficulty (p-value) as each item's mean response value;
2. Frequency table as response distribution across the 5 points scales;
3. Variability index as the standard deviation;
4. Biserial correlation between each item and the sum score with item removed; and

Table 5.3.
Correlation table of tested items in field test round 1, include all 3 trials.

	item(s)	1	2	3	4	5	6	7	8	9	10	11
1	**Fluidity**	1										
2	**Enthusiasm**	0.55	1									
3	**Curiosity**	0.58	0.73	1								
4	**Discovery**	0.42	0.57	0.58	1							
5	**Clarity**	0.58	0.41	0.61	0.55	1						
6	**Storytelling**	0.47	0.61	0.63	0.61	0.57	1					
7	**Creativity**	0.49	0.61	0.6	0.48	0.37	0.5	1				
8	**Entertainment**	0.57	0.74	0.82	0.57	0.54	0.7	0.59	1			
9	**Untroubling**	0.34	0.25	0.34	0.15	0.4	0.26	0.03	0.37	1		
10	**Captivation**	0.53	0.72	0.62	0.56	0.5	0.6	0.52	0.7	0.2	1	
11	**Pleasing**	0.74	0.62	0.73	0.49	0.72	0.63	0.52	0.65	0.33	0.63	1

5. Pattern of average total test scores (the mean column) for persons in each response category for each item.

Ideally, the items should demonstrate a moderate level of difficulty (around 3 for a 5-point Likert scale), and an adequate variability index (more than 1 is preferred). In addition, I would like to see an approximately normal distribution of responses across all scales, where a reasonable positive correlation between each item score and sum score of rest of the items existed (more than 0.7 is preferred). At the same time, a pattern where respondents who answer a lower score on an item will correspond to a lower total score in the category table is expected.

Figure 5.2 shows summary of 11 items in field test round 1. The colored squares show the label of each item. For each of item, the graph consists of the item response distribution (point 1 - point 5), the variability index, and polyserial correlation between each item and the sum score with item removed. By viewing the item response distribution plots, in general, all items shown an adequate level of item difficulty and

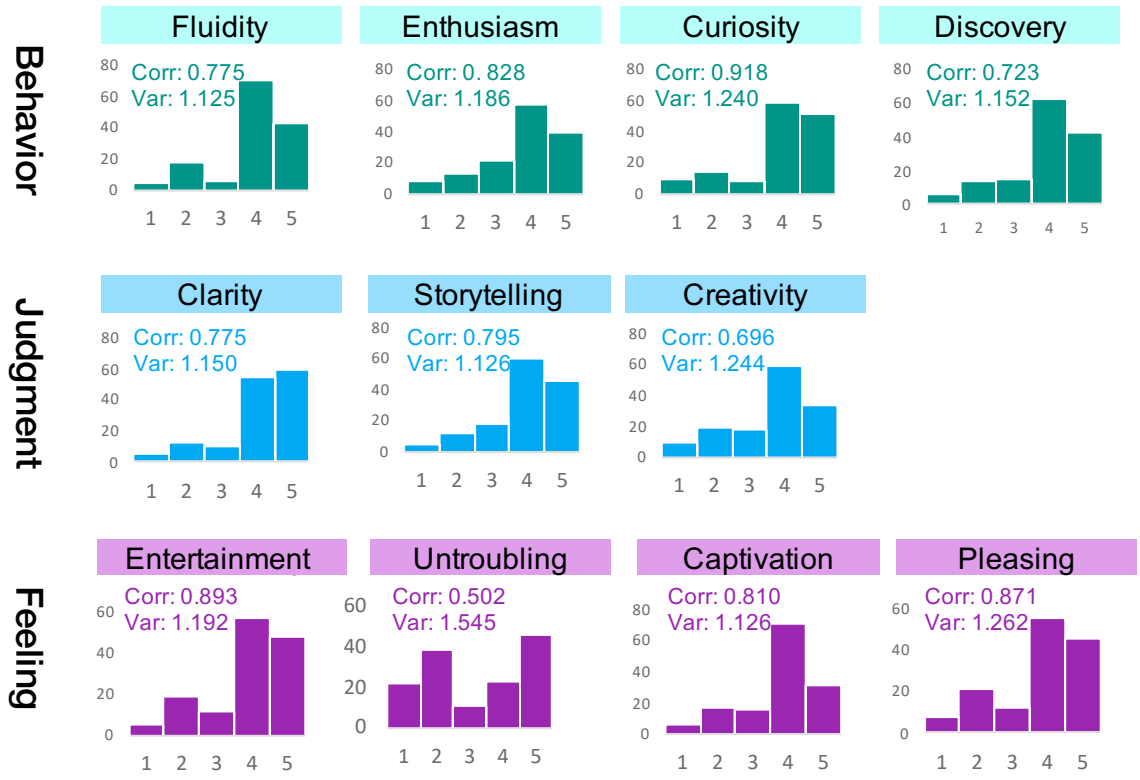


Figure 5.2. Summary of item analysis from 11 items in field test round 1. The order of the items in the instrument is from left to right and top to down, categorized by three levels (Behavior, Judgement, and Feeling).

variability index. However, while 10 of the items have roughly bell-shaped-like distributions with somewhat right skews, the Untroubling item showed an inverted bell curve (see "Untroubling" in the lower middle of Figure 5.2). As for the biserial correlation, the correlation between Untroubling item and the sum score is significantly lower than the rest of the items as well.

In summary, the response pattern of item "Untroubling" suggests that respondents were intentionally avoiding the central option (neither agree or disagree) in the scale, which implies issues in respondents' statement interpretation. On the other hand, the systemically skewed distributions of the item responses is understandable since

all three visualizations in the field test were well-designed and are award-winning visualizations.

5.3 Field Test Round 2

Based on the results from the first field test, the research team decided to revise Untroubling item due to its problematic response distribution (inverted bell shape) and low biserial correlation (around 0.5) compared to the other 10 items. Furthermore, a considerably high score across all three tested visualizations was observed, due to the fact that all of them are award-winning works from a contest. The team believe it would be more appropriate to test visualizations that are not that good, so that we can investigate how items would behave under different situations. Hence, the tested visualization no. 2 was replaced with a more mediocre visualization created by the research team (See Figure 5.3). This interactive chart maintained the same characteristics in terms of size and interactivity, and kept a similar topic of the original no.2, but was intentionally made as an average-quality work.

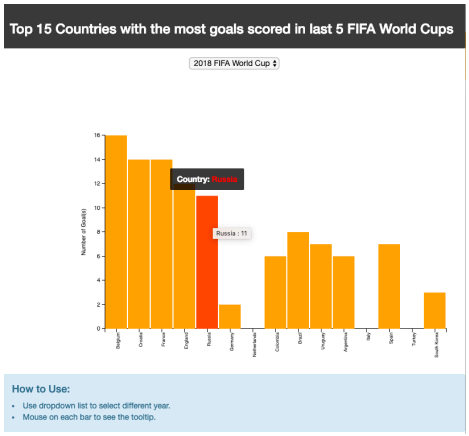


Figure 5.3. The modified tested visualization no.2. Created by the research team member Ali Baigelenov, with Ya-Hsin Hung's further editing.

After an internal discussion within research team, to make the revision process more efficient, besides modifying the original item, we chose two additional candidates

Table 5.4.
Average of total score, total time spent on trial, and total time spent on the instrument for the three tested visualisation in field test round 2, include all 3 trials, N = 54.

Candidate	**Statement**
v1	I feel **frustrated** when viewing this visualization.
v2	I feel no difficulty when viewing this visualization.
v3	I **don't** feel frustrated when using this visualization

from the item bank that were also able to assess indicator *Untroubling*, and conducted another round of field test with the 10 original items and the three newly added item candidates.

There are extensive discussions amd debates on whether to use negative wording on the Likert-scale items [193, 194]. Therefore, different strategies were utilized to improve the original Untroubling item. For the v1 candidate, the term "frustrated" in the original item was emphasized, making it underlined and bold, which is a common practice in a reverse-coded item. And for the v2 candidate, another candidate was chosen from the item finalist corresponding to indicator "untroubling". For the v3 candidate, I modified the first version and make it into a double-negative statement so that there is no need to reverse coded the responses. The three candidates items are shown in Table 5.4.

With the updated version of visualization 2, and a modified online survey form, this time 54 participants were recruited for the field test. After dropping responses from one participant who did not satisfy the filtering criteria, round 2 ended up with 53 participants providing a total of 159 survey responses. The average total score, the average time spent for the trial, and the time spent for filling the post-trial survey for three tested visualizations are shown in Table 5.5.

Following the same approach described in the previous section, item analysis was conducted again with the second round collected data. As shown in Figure 5.4, the

Table 5.5.
Average of total score, total time spent on trial, and total time spent on the instrument for the three tested visualization in field test round 2, include all 3 trials, N = 54.

	Viz no1	**Viz no2**	**Viz no3**
Total Score (point)	42.30	42.35	43.15
Time Spent on Trial (sec)	117.64	56.39	146.11
Time Spent on Instrument (sec)	36.60	24.81	40.58

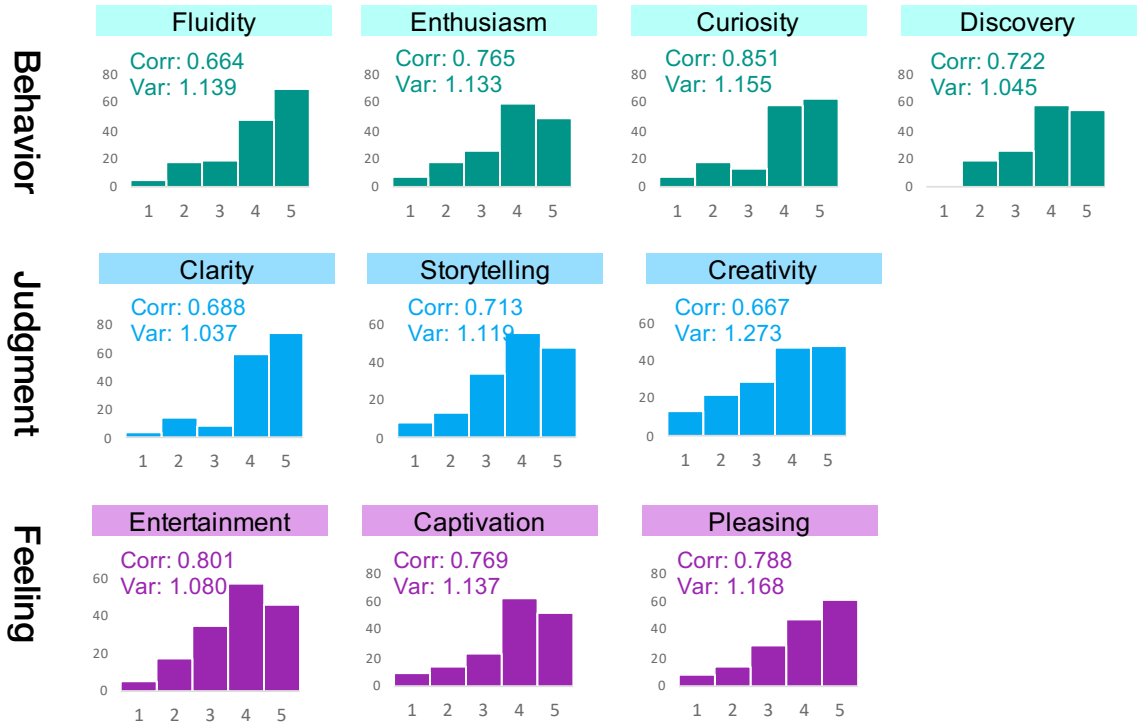


Figure 5.4. Summary of item analysis from 10 items except Item Untroubling in field test round 2. The order of the items in the instrument is from left to right and top to down, categorized by three levels (Behavior, Judgement, and Feeling).

original 10 items still performed decent by looking at the item analysis statistics. As for the three replacement candidates for item Untroubling, the results (see Figure

5.5) indicate that the 3rd candidate has the most preferable response patterns and the highest correlation with the sum score.

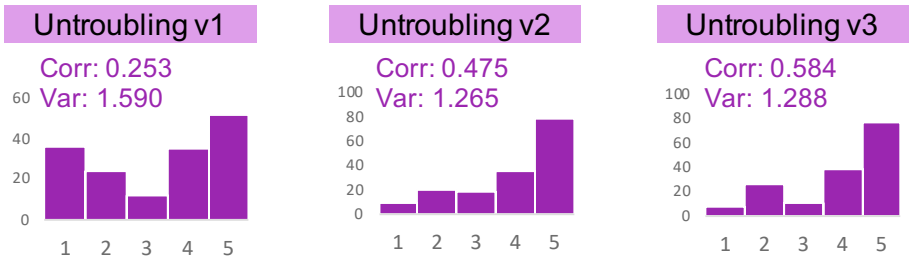


Figure 5.5. Summary of item analysis for 3 candidates of Item "Untroubling" in field test round 2.

Based on the item analyse results, the second version of AEVis with item v3 was assembled. Since it is believed that a double-negative statement using Likert-scale requires more attentional resources from respondents [195], the item sequence was re-ordered so that the item v3 was moved to the third place in the survey instrument. As a result, the updated item order in the second version of AEVis is: **Fluidity, Enthusiasm, Untroubling, Curiosity, Discovery, Clarity, Storytelling, Creativity, Entertainment, Captivation, and Pleasing**.

5.4 Field Test Round 3

In this round of data collection, responses from 221 participants were collected. Again, each participant provided survey responses on each of the three visualizations. With the same screening approach as described previously, 14 participants were dropped, leaving a total of 207 participants. Since the items tested here were also tested in the second round, I integrated the data of 53 participants from the previous stage, making a total of 241 participants for our data analysis.

Note that there were some modifications on the assigned tasks for tested visualization no.2 in the study. In the second round of field test, the time spent for the three trials are 117.64 (sec), 56.39 (sec), and 146.11 (sec), respectively. It is obvious that the time spent for the updated visualization no.2 trial is significantly lower than the others, which might suggest the content-related questions are too easy for the participants. Therefore, the research team modified the content-related questions in the third round of field test. This time, the average total score, the average time spent for the trial, the time spent for filling the post-trial survey for three tested visualizations are shown in Table 5.6.

Table 5.6.
Average of total score, total time spent on trial, and total time spent on the instrument for the three tested visualisation in field test round 3, include all 3 trials, N = 241.

	Viz no1	**Viz no2**	**Viz no3**
Total Score (point)	40.15	41.24	40.66
Time Spent on Trial (sec)	145.5	114.91	194.13
Time Spent on Instrument (sec)	37.10	35.07	34.43

Table 5.7.
Correlation table of tested items in field test round 3, include all 3 trials.

	item(s)	1	2	3	4	5	6	7	8	9	10	11
1	**Fluidity**	1										
2	**Enthusiasm**	0.52	1									
3	**Untroubling v3**	0.66	0.44	1								
4	**Curiosity**	0.59	0.59	0.5	1							
5	**Discovery**	0.38	0.47	0.29	0.49	1						
6	**Clarity**	0.55	0.45	0.52	0.54	0.35	1					
7	**Storytelling**	0.35	0.43	0.28	0.51	0.51	0.36	1				
8	**Creativity**	0.32	0.51	0.24	0.55	0.51	0.34	0.54	1			
9	**Entertainment**	0.42	0.55	0.36	0.68	0.54	0.42	0.59	0.61	1		
10	**Captivation**	0.43	0.56	0.41	0.6	0.51	0.41	0.54	0.63	0.63	1	
11	**Pleasing**	0.63	0.59	0.51	0.66	0.43	0.51	0.48	0.46	0.58	0.54	1

5.4.1 Factor Analysis

Factor analysis is a widely utilized and broadly applied statistical technique for describing variability of observed variables with respect to the unobserved latent variables, which can both test measurement integrity and guide further theory refinement [196, 197]. For instrument development, factor analysis models can be used to interpret whether there is a set of underlying latent traits with respect to the observed variables, therefore been widely used in social science [198]. Information about the number and structure of latent traits (i.e., factors, constructs) underlying a set of test or item scores can provide evidence for validation of particular test score interpretations. By investigating factor loading and distribution of each factor across the items, factor analysis can be utilized to test the internal structure and external structure of the instrument [180].

Results from factor analysis can also be helpful in making decisions on whether to remove or revise certain instrument items [31]. Several sets of factor loadings can be found on each item from factor analysis, where factor loading refers to the relationship of each variable to the underlying factor. A higher factor loading implies that there are higher correlations between the item and a "set". Based on the loadings, items can be grouped into several clusters (a loading of 0.4 or greater is preferred) and fitted into several candidate factor models. Information about the number and structure of latent traits underlying a set of test or item scores can provide evidence for validation of particular test score interpretations.

Several cutoff for criteria can be considered when judging the number of meaningful factors [199]: (1) substantive interpretability of the loading and factor correlation pattern, (2) model fit, and (3) parsimony. Overall, Model fit statistics can be interpreted by: Chi-square test of model fit (prefer p-value >0.05), Tucker-Lewis index (TLI, prefer to exceed 0.95), and root mean square error of approximation (RMSEA, prefer <0.06).

Furthermore, reliability coefficient omega [200] (prefer to be>0.7) can also be helpful, where reliability is a property of observed test scores from a particular instrument in a specified examinee population. There are extensive discussions over the usage of various measures of reliability of the total score (e.g., coefficients alpha, beta, and omega) [201, 202]. Even though coefficient alpha has been widely adopted [203], since it is known that coefficient alpha underestimates the true reliability unless the items are tau-equivalent, the coefficient omega was selected as a practical alternative in this study [204].

5.4.2 Analysis

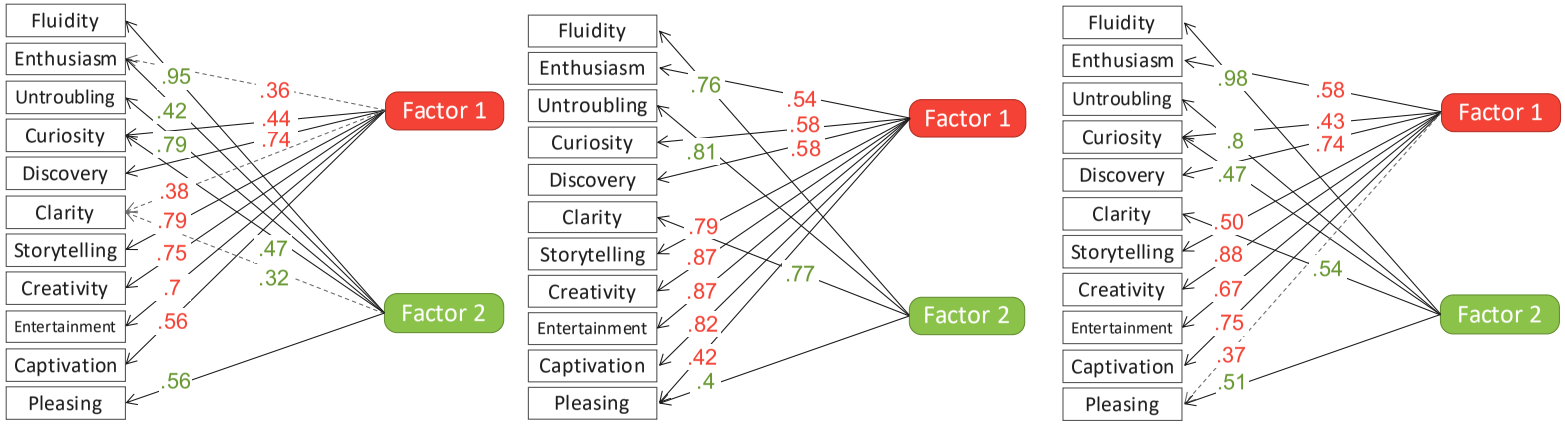
The exploratory factor analysis (EFA) approach was adapted to analyze the data we collected in the third round of field test. The primary purpose of EFA is to explore the appropriate number of underlying factors which would fit the dataset (responses

from three tested visualization). EFA is believed to be a useful tool that is able to aid the researchers in recovering an underlying measurement model, and then can be evaluated with CFA [185, 205]. I utilized R packages psych [191] and GPARotation [206] to conduct EFA with four different settings (number of underlying latent factor(s) = 1, 2, 3, and 4), using Pearson correlation and oblimin rotation. Typically, a range of n (number of underlying latent factor) would be tested throughout the EFA process. Due to the fact that a three-level model was proposed in the stage 1 (behavior, judgment, and feeling), I used a convenience cut-off n=4 as the stop point for testing the number of underlying latent factor in EFA.

Several model fit statistics were checked when reviewing the EFA results and judged the model fit, which include: Chi square, TLI, and RMSEA. Generally, we want to find the simplest model with an acceptable level of model fitness. That is, we would like to choose models with lower number of factors but still have acceptable model fit statistics. The four types of model fit statistics mentioned above under 4 different conditions (number of factors = 1, 2, 3, and 4) were listed in the three tables, grouped by the tested visualization (see Table 5.8, 5.9, and 5.10).

Table 5.8.
Model Fit Statistics of tested visualisation no. 1 dataset under 4 different conditions (number of factors = 1, 2, 3, and 4).

Visualisation no. 1	**n=1**	**n=2**	**n=3**	**n=4**
Likelihood Chi Square	134.33	60.21	41.65	22.69
Tucker Lewis Index (TLI)	0.92	0.97	0.974	0.987
Root Mean Square Error of Approximation (RMSEA)	0.094	0.032	0.021	0.04
Reliability coefficient Omega (Omega)	0.94	0.93	0.95	0.93



(a) EFA results of tested visualizations no.1 [169] when n=2.

(b) EFA results of tested visualizations no.2 (modified) when n=2.

(c) EFA results of tested visualizations no.3 [171] when n=2.

Figure 5.6. EFA results of three tested visualizations when n=2. Each square in the path diagram represents one item, Factor 1 and Factor 2 on the right-hand side represent two underlying latent factors.

Table 5.9.
Model Fit Statistics of tested visualisation no. 2 dataset under 4 different conditions (number of factors = 1, 2, 3, and 4).

Visualisation no. 2	**n=1**	**n=2**	**n=3**	**n=4**
Likelihood Chi Square	270.54	39.65	16.44	9.53
Tucker Lewis Index (TLI)	0.761	0.992	1	1
Root Mean Square Error of Approximation (RMSEA)	0.13	0.029	0	0
Reliability coefficient Omega (Omega)	0.93	0.91	0.94	0.92

Table 5.10.
Model Fit Statistics of tested visualisation no. 3 dataset under 4 different conditions (number of factors = 1, 2, 3, and 4).

Visualisation no. 3	**n=1**	**n=2**	**n=3**	**n=4**
Likelihood Chi Square	73.43	94.77	62.85	37.37
Tucker Lewis Index (TLI)	0.906	0.943	0.952	0.962
Root Mean Square Error of Approximation (RMSEA)	0.112	0.088	0.081	0.073
Reliability coefficient Omega (Omega)	0.95	0.94	0.96	0.94

5.4.3 Findings

After investigating different number of factors using EFA approach, by reviewing the selected fit statistics, I determined the most plausible underlying structure for our collected data. For all three tested visualizations, the 2 factor models were the most appropriate, although the pattern of item loadings on factors was not identical. The reliability coefficient (omega) for each of them are 0.93, 0.91, 0.94, respectively, which suggest a decent level of response reliability (more than 0.9). Then, the interpretive descriptions of the selected two factor model were made.

As shown in Figure 5.11, the allocation of items is similar across 3 tested visualizations. items that placed within parentheses means the loadings among factor 1 and factor 2 are close and therefore will show up under both factors.

1. The EFA results of tested visualization no.1 (creative routine) [169] when n = 2 shows allocation of items as below:
   - Factor 1: (Enthusiasm), (Curiosity), Discovery, (Clarity), Storytelling, Creativity, Entertainment, Captivation
   - Factor 2: Fluidity, (Enthusiasm), Untroubling, (Curiosity), (Clarity), Pleasing
2. The EFA results of tested visualization no. 2 (FIFA World Cup goals) when n = 2 shows allocation of items as below:
   - Factor 1: Enthusiasm, Curiosity, Discovery, Storytelling, Creativity, Entertainment, Captivation, (Pleasing)
   - Factor 2: Fluidity, Untroubling, Clarity, (Pleasing)
3. The EFA results of tested visualization no. 2 (FIFA World Cup goals) when n = 2 shows allocation of items as below:
   - Factor 1: Enthusiasm, (Curiosity), Discovery, Storytelling, Creativity, Entertainment, Captivation, (Pleasing)
   - Factor 2: Fluidity, Untroubling, (Curiosity), Clarity, (Pleasing)

In summary, item Enthusiasm, Discovery, Storytelling, Creativity, Entertainment, and Captivation have larger loadings (>0.4) on latent factor 1, and therefore indicated stronger associations with this factor. On the other hand, item Fluidity, Curiosity, Clarity, Untroubling, and Pleasing show stronger loadings on latent factor 2. Based on the assortment of the indicators under each factor, a substantive interpretation of the loading and factor correlation pattern is revealed:

- Factor 1 consists of items related to Enthusiasm, Discovery, Storytelling, Creativity, Entertainment, and Captivation–I therefore interpreted that this latent factor might be **Hedonic** aspect of AE.

- Factor 2 consists of items related to Fluidity, Curiosity, Clarity, and Untroubling– I then interpreted this latent factor might be the **Pragmatic** aspect of AE.

Table 5.11.
Factor loadaings of factor 1 and factor 2 from three tested visualizations under situation of 2-factor model.

Item	**Visualization no.1**		**Visualization no.2**		**Visualization no.3**	
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2
Fluidity	-0.17	0.95	0.05	0.76	-0.12	0.98
Enthusiasm	0.36	0.42	0.54	0.19	0.58	0.22
Untroubling	-0.08	0.79	-0.13	0.81	-0.05	0.8
Curiosity	0.44	0.47	0.58	0.27	0.43	0.47
Discovery	0.74	-0.1	0.58	0.11	0.74	-0.06
Clarity	0.38	0.32	-0.11	0.77	0.24	0.54
Storytelling	0.79	-0.13	0.79	-0.05	0.5	0.21
Creativity	0.75	0.01	0.87	-0.19	0.88	-0.13
Entertainment	0.7	0.16	0.87	-0.1	0.67	0.17
Captivation	0.56	0.25	0.82	-0.13	0.75	0.1
Pleasing	0.27	0.56	0.42	0.4	0.37	0.51

During stage 2, several revisions on the experiment materials and test beds have been made throughout the field test. For example, the change of tested visualization no.2 and the modifications on the content-related questions. Interestingly, a significant lower score from the updated visualization no.2 was not observed as expected, however, it is likely that the content of the visualization (FIFA World Cup) plays an role here. It is possible that information regarding world cup is still relatively popular than the topics of other two tested visualizations, and therefore gaining more scores in the items such as “Entertainment” and “storytelling” This results inline with our theory that AE is a complex construct and could be influenced by multiple factors in the stage 1.

It is worth to note that, an existing instrument for evaluating perceived usability (user experience) UEQ [207] for which similar subscales were identified empirically, also identified two underlying factors as hedonic vs. pragmatic (More discussions regarding other similar established instruments can be found in discussion section). Due to the fact that by definition, like user experience, AE is also a subjective human trait (experience) comes from what users perceived mentally and physically from the visualization they interacted with. A similar results from an existing study is able to provide additional validation support for the claim that this factor structure is likely to be replicable in the population other than the one in the field test. Additionally, there are similarities between the proposed 2-factor model and the alternative explanation of AE construct been briefly mentioned in the stage 1 findings. The conceptual space as "person trait" vs. "interaction between person and visualization" is to some extent comparable to the hedonic vs. pragmatic structure; while person trait is more related with hedonic aspect, the consequences between person and visualization is more related to pragmatic aspect.

Finally, Table 5.1 listed all items been tested in the field tests, and its factor loadings of two latent factors Hedonic and Pragmatic when responses of all three tested visualizations fitted in an EFA model. The results of EFA indicates that even with three different testbeds, participants' response is considerably stable. Moreover, the allocation of items from 2-factor models can make reasonable interpretation of underlying latent factors that are not too far-stretched. Therefore, the EFA results support the validity evidence of internal structure of the instrument.

5.5 Comparisons between Three Level and Two Factor Models

CFA (Confirmatory Factor Analysis) is often used during the process of instrument development to examine the latent structure of a test instrument. And is a confirmatory technique where a hypothesized model is specified to estimate a population covariance matrix that is compared with the observed covariancematrix [208,209]. In CFA, the number of factors required in the data would be specified, and researchers can study the relationships between observable measured variable and latent variable that they are interested in.

Based on the analysis in the field test, the results from Exploratory Factor Analysis (EFA) suggest a two-factor model which contains two underlying factors of our developed AEVis instrument: pragmatic and hedonic aspects. However, our original theory in stage 2 proposes a different three-level structure of AEVis: behavior, judgment, and feeling. In practice, it is quite common that results of factor analysis are different from the original theoretical background (See [208], Chapter 5). Because respondents' response patterns and the observed correlations among items often deviate from those expected under the theoretical framework or functional grouping predicted by the researcher. Therefore, to determine whether a certain factor solution is a better representation of the data than the other factor solutions, in this section, a model comparison using CFA approach was conducted to further investigate the differences between the two models and their potential explanations.

5.5.1 Method

Since the purpose of this study is to determine which model can better represent the data out of the two proposed model, the analyzed data is identical to the dataset used in the EFA which contains 723 participants' responses. In order to conduct CFA, R packages lavaan [210] and Semplot [211] were utilized. The dataset is fitted into two models: 3-level and 2-factor models.

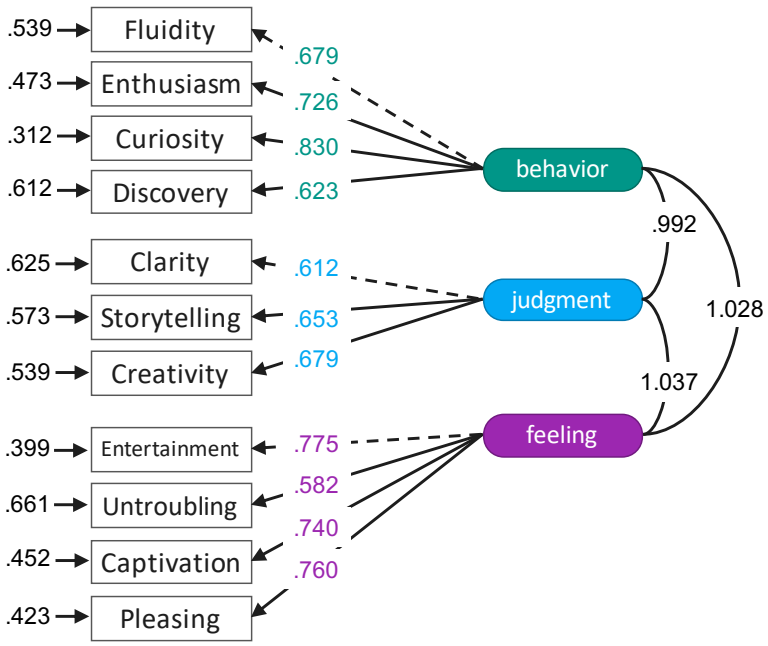


Figure 5.7. CFA path diagram of the 3-level model, 3 circles represent 3 underlying latent variables defined in the model: behavior, judgement, and feeling; colored numbers represent the standardized factor loadings for the items.

- The structure of the **3-level model** is denoted below: item Fluidity, Enthusiasm, Curiosity, and Discovery are assigned under "behaviour" factor; item Clarity, Storytelling, and Creativity are assigned under "judgment" factor; finally, item Entertainment, Untroubling, Captivation, Pleasing are assigned under "feeling" factor (See Figure 5.7).
- The **2-factor model** is denoted as below: item Fluidity, Curiosity, Clarity, Untroubling, and Pleasing are under "pragmatic" factor; then, item Enthusiasm, Discovery, Storytelling, Creativity, Entertainment, and Captivation are assigned under "hedonic" factor (See Figure 5.8).

The results of CFA can be plotted into path diagrams. Each square represent an observable variable (item), the arrows between circles and squares indicate the relationship that we defined in the model, and the numbers represent the standardized

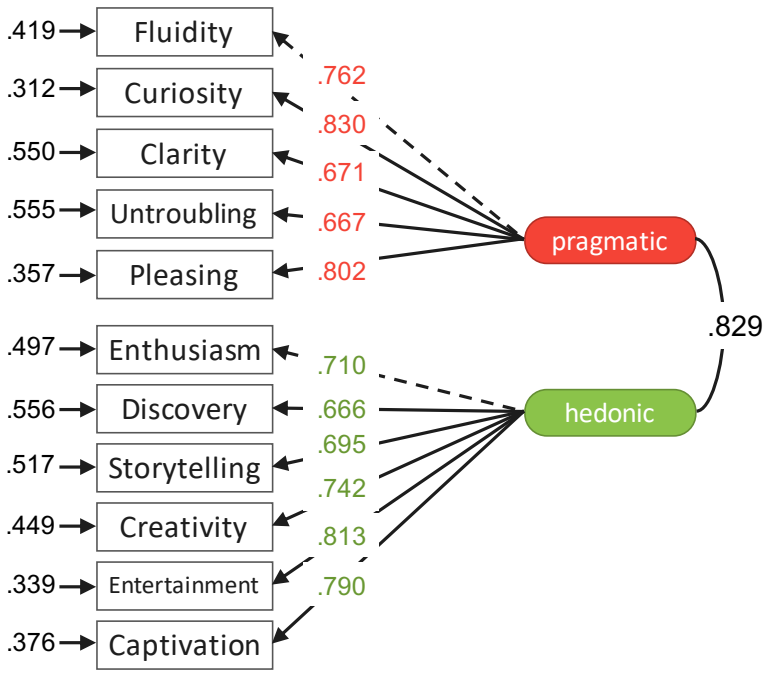


Figure 5.8. CFA path diagram of the 2-factor model, 2 circles represent 2 latent variables defined in the model: pragmatic and hedonic; Colored numbers represent the standardized factor loadings for the items.

factor loadings. Arrows between circles are standardized covariances between latent variables. Finally, the bottom part of the numbers represent the standardized variance of each item.

As shown in Figure 5.7 and 5.8, while factor loadings on the items and latent variables are generally preferred to be >0.7, both models have adequate level of factor loadings. Therefore, We will look at several model fit statistics (see Table 5.12) to help us make decision on which model is more appropriate.

Generally, less error implies that data is close to model, which means this is a more accurate representation of the relationship that people trying to model. As Schreiber summarized the cut-off criteria for several fit index in a review of CFA and related statistical analytical techniques [209]. For CFI and TLI, values that >0.95 are preferred. As for AIC and BIC, models have lower values on these two indices are

Table 5.12.
Model Fit statistic indice of 2 tested models, the first column contains 6 model fit indices examined in this study.

	3 level model	**2 factor model**
Comparative Fit Index (CFI)	0.875	0.926
Tucker-Lewis Index (TLI)	0.832	0.905
Akaike information criterion (AIC)	21217.281	20993.919
Bayesian information criterion (BIC)	21331.866	21099.337
Root Mean Square Error of Approximation (RMSEA)	0.135	0.102
Standardized Root Mean Square Residual (SRMR)	0.071	0.057

preferred. As for RMSEA and SRMR, since they are residuals of the model, the rule of thumb is to pick values that are <0.06 to 0.08 and <0.08, respectively.

5.5.2 Findings

As shown in Table 5.12, both model fits are acceptable but not excellent. But all 6 indices indicate that the 2-factor model has a slightly better model fitness than the 3-level model. In practice, when interpreting the models, people tend to consider parsimony as an important characteristic of the model which suggests that a smaller number of factor should be preferred. Additionally, by looking at the theoretical definitions of two models and their latent variables, while the 3-level structure can surely be identified, there is no strong conflict against the meta-dimension of pragmatic vs. hedonic aspects.

These results, with findings from stage 1, confirm that "engagement" is a complex and multifaceted construct. It is very likely that both 3-level and 2-factor structures are co-existed in the construct space, but the 2-factor structure is more impactful and can be better explained by its model. Thus, based on the CFA results in this section,

I would primary suggest a 2-factor model for AEVis, but still adapt the 3-level theory as the secondary way to explain the construct.

5.6 Re-examine with Item Response Theory

Item Response Theory (IRT) models such as Rasch model, nominal response model, graded response model are able to show the relationship between the ability or trait (target construct) measured by the instrument and an item response [212]. By using IRT, researchers can retrieve more information about items characteristics (e.g., discrimination ability, test information function) compared to traditional CTT (classical test theory) methods.

Based on current intended uses and characteristics of target construct (i.e., 2 dimensionality, conditional independence), the 2PL multi-dimensional IRT model for polytomous data was picked as our modeling method. Since the response categories are assumed to be ordered (i.e. Likert scale from strongly agree - 5 to strongly disagree - 1), the Graded Response Model is preferred.

Like factor analysis, several criteria can be considered when judging model fit: Akaike information criterion (AIC, the lower the better model fit [213]); Bayesian information criterion (BIC, the lower the better model fit [214]); and Root mean square error of approximation (RMSEA, prefer <0.05). With IRT modeling technique, we can retrieve several item parameter estimations that can be utilized to identify potential issues on items or scales such as Discrimination Parameters (a) and Threshold Parameters (b, similar to item difficulty).

5.6.1 Method

Adopting an identical dataset from EFA phrase which contains 723 responses from field test round 2 and 3. This time, IRT-PRO [215] software version 4.1 was utilized to analyze data, with Bock-Aitkin EM Algorithm for estimation. We analyzed the

Table 5.13.
Comparisons between four IRT models with collected data in field test round 3. For each model, discrimination parameters (a) as well as three model fit statistics (AIC, BIC, and RMSEA) are listed.

	2Dim GRM (Explore)		**2Dim GRM (ESEM)**		**2Dim GRM (Confirm)**		**UniDim GRM**
Item	a_1	a_2	a_1	a_2	a_1	a_2	a
Fluidity	3.66	0.76	3.68	0	0	2.56	1.95
Enthusiasm	1.59	1.48	0.97	1.42	2.07	0	2.19
Untroubling	2.58	0.36	2.45	0	0	1.89	1.5
Curiosity	2.35	2.17	1.44	2.09	0	3.24	3.28
Discovery	0.91	1.58	0	1.83	1.86	0	1.7
Clarity	1.75	0.72	1.93	0	0	1.86	1.61
Storytelling	0.89	1.95	0	2.12	2.07	0	1.83
Creativity	0.87	2.49	0	2.41	2.32	0	1.94
Entertainment	1.57	2.85	0	3.34	3.09	0	2.74
Captivation	1.42	2.28	0	2.7	2.76	0	2.44
Pleasing	2.32	1.61	1.75	1.38	0	2.98	2.66
AIC	18448.66		18473.57		18685.22		18892.88
BIC	18751.17		18743.99		18941.89		19154.13
RMSEA	0.08		0.08		0.08		0.08

collected data with four different models, ordered by the most restricted to the least restricted:

- **UniDim GRM:** uni-dimensional graded response model;
- **2Dim GRM-Confirm:** 2-dimensional graded response model with full constraint, all items are assigned to one of the two dimensions, similar to CFA;

- **2Dim GRM-ESEM:** 2-dimensional graded response model with partial constraint (Exploratory Structural Equation Model), the factor assignment is based on EFA results from previous section. Some of the items have been assigned to one of the dimensions, while some of them have no assignment; and
- **2Dim GRM-Explore:** 2-dimensional graded response model without constraint, no assignment for both dimensions, similar to EFA.

The item discrimination parameter estimates (a) can be found in column "a_1" and "a_2" in Table 5.9, and item threshold parameters estimates (b) can be founded in "b_1", "b_2", "b_3", and "b_4" of Figure 5.10. The item discrimination parameters in IRT models are conceptually the same as the factor loadings from linear FA. Compared to linear FA, nonlinear FA models such as IRT can produce a test information function (information curve in uni-dimensional IRT model), and indicate the difficulty of specific response categories for each item.

Figure 5.9 and 5.10 are visual representations of threshold parameters of 11 items in two of the fitted 2-dimensional graded response models – 2Dim GRM-ESEM and 2Dim GRM-Explore. For each item, the four threshold parameters represent the boundaries between the five response options on a conceptual spectrum of latent trait θ (e.g., b_1 means the boundary between point 5 and point 4, b_2 means the boundary between point 4 and point 3, so on so forth).

5.6.2 Findings

As shown in Figure 5.13, 2-dimensional graded response model with partial constraint (2Dim GRM-ESEM) and 2-dimensional graded response model without constraint (2Dim GRM-Explore) have lower AIC and BIC which suggested better model fit. While all four models have RMSEA at 0.08, this indicate not supreme, but adequate model fit for all 4 models.

Overall, all models show adequate level of discrimination parameters across items and the assigned factors. Even though item Enthusiasm ($a_2 = 0.97$) in 2Dim GRM

(pre-EFA) model has the lowest a, it is still significantly higher than 0.4 (the minimum preferred value of a).

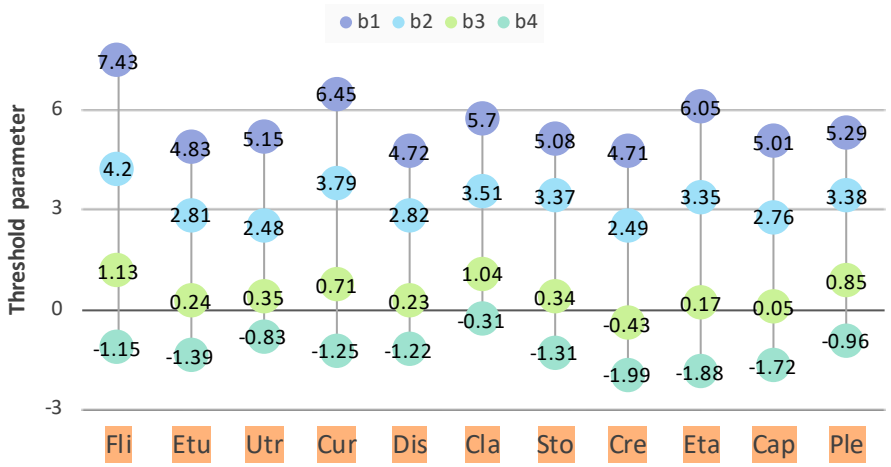


Figure 5.9. Threshold parameters of 11 items in the fitted 2-dimensional graded response model (2Dim GRM-Explore), with item labels listed at the bottom.

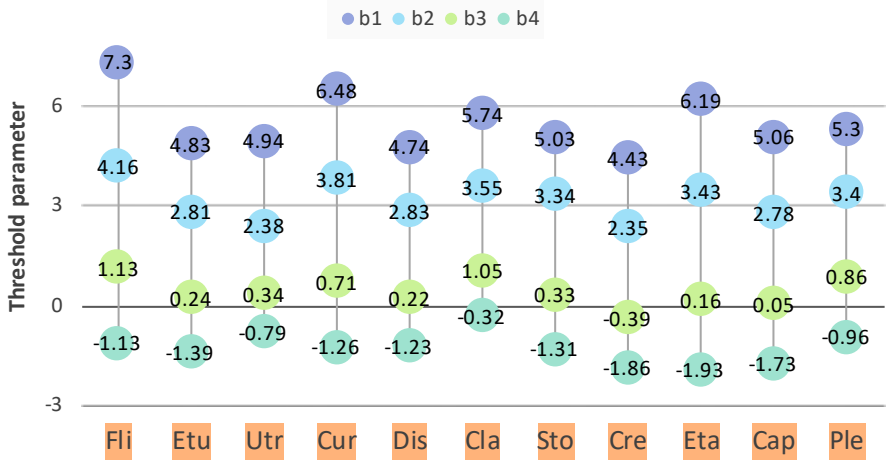


Figure 5.10. Threshold parameters of 11 items in the fitted 2-dimensional graded response model (2Dim GRM-ESEM), with item labels listed at the bottom.

As for threshold parameters of 2Dim GRM-ESEM and 2Dim GRM-Explore models, the distances between threshold parameters seems adequate, and boundary spacing appears fairly uniform as well, as shown in both Figure 5.9 and Figure 5.10). It indicates that 11 items appear to have reasonable capacity to differentiate among respondents with different latent trait θ status (AE of the respondents). Although distance between b_3 and b_4 in item Untroubling is a potential concern. Generally, no item shows significantly low distance between each threshold parameter or overlapping threshold parameters, which suggests the instrument scale is at an acceptable level and will presumably function normally across different populations other than the tested one.

By utilizing IRT models, an exploratory approach using a confirmatory model was adopted, which is able to provide additional evidence to support the 2-factor structure from EFA results. The results of IRT analysis to some extent re-confirmed results from item analysis and FA. The 2-factor model which consists of hedonic and pragmatic aspects of AE indicates a decent model fitness with the collected data from field test. Finally, by examining discrimination parameters and threshold parameters of the two most fitted models (i.e. 2Dim GRM-ESEM and 2Dim GRM-Explore), the developed items demonstrate a decent level of difficulty and discrimination ability.

In summary, by testing AEVis on three visualizations in the field tests, the underlying construct of developed instrument and how its items behaved with different data, was investigated. Results from the three analytical methods (item analysis, FA, and IRT) contribute to instrument revisions and score interpretation validations.

6. STAGE 4–EVALUATION OF AEVIS

In this chapter, a follow-up field test was conducted to investigate the external relation of AEVis and other related instruments. It is a common practice in instrument development to test the correlations between target instrument and others [156]. This was done by experimenting visualization users' level of affective engagement and level of other factors via multiple surveys and questionnaires, and then studying the correlations between the collected scores (IRB protocol#: 1903021883).

The purpose of this type of field test is to find whether there are relations between known scores. For example, for an IQ test (Intelligence Quotient), the score interpretations are meant to have positive correlations with standard test scores such as GRE (Graduate Record Examination) or GPA (Grade Point Average), and if the tested students' IQ test score shows positive correlations with his or her SAT score as well as GPA from multiple semesters. Then, one can say the score interpretations of this IQ test are supported by the external relation evidences.

The goal of this study is to:

1. Evaluate and collect validity evidence of the affective engagement survey instrument that developed in this dissertation, and;
2. Learn how affective engagement, as a construct, relates or interacts with other user experience factors such as perceived usability and user engagement.

6.1 External Relations between AEVis and Other Instruments

6.1.1 AEVis and UEQ-s

The first external relations that will be investigated are between AEVis (Affective Engagement Visualization Survey) and UEQ-s (User Experience Questionnaire-short)

instruments. Both AEVis and UEQ-s assume two underlying factors in their instruments. For AEVis, they are pragmatic aspects (5 items) and hedonic aspects (6 items) of affective engagement. As for UEQ-s, the two dimensions are pragmatic quality (4 items) and hedonic quality (4 items) of perceived usability [207].

The description of the 5 factors in UEQ-s are listed below, their corresponding items as well as factors can be found in Table 6.2:

- **Perspicuity**: Is it easy to get familiar with the product? Is it easy to learn? Is the product easy to understand and clear?
- **Efficiency**: Can users solve their tasks without unnecessary effort? Is the interaction efficient and fast? Does the product react fast to user input?
- **Dependability**: Does the user feel in control of the interaction? Can he or she predict the system behavior? Does the user feel safe when working with the product?
- **Stimulation**: Is it exciting and motivating to use the product? Is it fun to use?
- **Novelty**: Is the product innovative and creative? Does it capture users attention?

These two instruments cover different elements in terms of languages. For example, the AEVis pragmatic aspects includes items for Fluidity, Curiosity, Clarity, Untroubling, and Pleasing. In contrast, UEQ-s pragmatic quality includes items for Dependability, Perspicuity, and Efficiency. Still, the meta-dimensions of pragmatic vs. hedonic in two instruments are theoretically similar, and therefore, they should demonstrate strong or moderate positive correlations between comparable dimensions of two instruments. In other words, the total score of items under AEVis pragmatic aspect should have a strong or moderate positive correlation with total score of items under UEQ-s pragmatic quality. The same should be true for hedonic aspects and hedonic quality in the two instruments as well. Finally, due to the theoretically similar internal structure, even though the differences between the two instruments will be more significant when we compare the entire instruments instead of subsets of them,

we still expect to see moderate or marginal positive correlations between the total scores of AEVis and UEQ-s.

6.1.2 AEVis and UES (subset)

The second external relations that will be investigated are between AEVis (Affective Engagement Visualization Survey) and a subset items of UES (User Engagement Scale) instruments. UES is a self-report measure that can assess user engagement with information systems, specifically oriented toward online news website. There are 31 items and total of 6 factors: aesthetic appeal, perceived usability, felt involvement, novelty, focused attention, and endurability [12].

The description of the 2 selected factors in UES are listed below, their corresponding items as well as factors can be found in Table 6.3:

- **Felt involvement**: The Felt Involvement factor contained items about how much fun users' were having during the interaction and how drawn in they were able to become.
- **Novelty**: Novelty in online content has the potential to sustain users' attention, specifically when novelty is introduced through links and content that are pertinent to users' goals.

After reviewing the descriptions of the 6 factors, potential similarities between some of the factors in UES and AEVis were identified. First, Felt involvement (3 items) in UES has resemblance to Enthusiasm (1 item) and Captivation (1 item) in AEVis on users fun and drawn experience. Second, Novelty (3 items) in UES echos with Entertainment (1 item) and Curiosity (1 item) in AEVis as both of them represent users curiosity and interests during the interaction. Therefore, the total scores of the above factor pairings were expected to show a moderate to marginal positive correlations with each other. To be specific, the total score of 3 items under Felt involvement in UES and the total score of 2 items under Enthusiasm (1 item) and Captivation (1 item) in AEVis should have marginal positive or moderate correlations.

The total score of 3 items under Novelty in UES and the total score of 2 items under Entertainment (1 item) and Curiosity (1 item) in AEVis should have marginal or moderate positive correlations.

6.2 Methods

The same as previous field tests, participants were recruited through Amazon Mechanical Turk platform. The instructions for the experiment were provided on the website, participants who were willing to participate then were redirected to the Qualtrics online survey system. Before participants start any research activities, the online consent form was shown and the participants had to agree to the consent form by clicking the check box in order to proceed further. When participants finished the experiment, they will be approved by the researchers and will automatically get paid $ 2.00 through the Amazon Mechanical Turk service.

Three surveys or subsets of a survey instrument were tested in the experiments, the time spent for each trial should be significantly longer than the previous field test. Hence, there were only 2 trials in the external relation study. Considering the content of UEQ and UES items, tested visualization no.2 and no.3 are appropriate. As a results, there are 2 trials in the experiment, 3 demographic questions will be asked at the end. The order of the two trials are randomized.

In each trial of this experiment, participants needed to:

1. Explore the provided interactive visualization;
2. Correctly answer 4 content-related questions of the provided interactive chart (one of the question is inserted in the post-trial surveys as an attention check);
3. Complete 3 post-trial surveys;
    - AEVis, 11 items
    - UEQ-s, 8 items
    - UES (subset), 6 items

6.2.1 Tested Survey Instruments

In this study, the full version of AEVis was used. There are 11 items in it, for each of them, respondents had to provide their degree of agreement upon the item statement (see Table 6.1).

Table 6.1.
11 tested items for AEVis, Dimensions and corresponding indicators are specified

Dimension	**Indicator**	**Item Statement**
Pragmatic	Fluidity	My use of this visualization is continuous and smooth.
Hedonic	Enthusiasm	I feel motivated while using this visualization
Pragmatic	Curiosity	I enjoy exploring this visualization
Hedonic	Discovery	I have acquired a new concept or new knowledge from the visualization
Pragmatic	Clarity	I think the visualization effectively delivers its main concept or idea
Hedonic	Storytelling	I think this visualization is telling a compelling story
Hedonic	Creativity	I think this visualization sparks my creative thinking
Hedonic	Entertainment	I feel entertained when using this visualization
Pragmatic	Untroubling	I don't feel frustrated when using this visualization
Hedonic	Captivation	I feel absorbed by this visualization while using it
Pragmatic	Pleasing	The look and feel of the visualization is pleasing to me

For UEQ-s, the full version which contains 8 items was used. For each of them, respondents had to choose a number between the conflicting terms better describes the product in the trial (see Table 6.2).

For UES, only a subset from its 33 items in the full version was used. There are 3 items for Felt Involvement and 3 items for Novelty, in total, 6 items for the UES was selected (see Table 6.3). Since UES was originally developed for news websites, some items clearly indicate the media or a specify user task. To adapt the tested visualizations used in this study, terms that are associated with news website and news reading were replaced (e.g., change "news website" to "visualization", change "visiting this website" to "using this visualization").

Table 6.2.
8 tested items for UEQ-s, dimensions and corresponding factors are specified.

Dimension	**Factor**	**Item**		
Pragmatic	Dependability	obstructive	o o o o o o o	supportive
Pragmatic	Perspicuity	complicated	o o o o o o o	easy
Pragmatic	Efficiency	inefficient	o o o o o o o	efficient
Pragmatic	Perspicuity	confusing	o o o o o o o	clear
Hedonic	Simulation	boring	o o o o o o o	exiting
Hedonic	Simulation	not interesting	o o o o o o o	interesting
Hedonic	Novelty	conventional	o o o o o o o	inventive
Hedonic	Novelty	usual	o o o o o o o	edge

Table 6.3.
6 tested items for UES, the corresponding factors are specified

Factor	**Item Statement**
Felt involvement 1	I was really drawn into finding the stories.
Felt involvement 2	I felt involved in this task.
Felt involvement 3	This experience was fun.
Novelty 1	I continued to read on the visualization out of curiosity.
Novelty 2	The content of the visualization incited my curiosity.
Novelty 3	I felt interested in the visualization.

6.3 Analysis

There were 48 participants in this study, their average age is 35.15 years old (Min 23, Max 62), and 18 of them are females. Responses that have significantly shorter

time spent (<10 sec) on both UEQ-s and UES were dropped. At the end, there were 40 valid responses for the analysis.

SPSS [216] version 25 was used to conduct correlation analysis. Since all three instruments adapt Likert-scale responses (ordinal data), Spearman correlation was used in the analysis which is often used to evaluate relationships involving ordinal variables. For each of the tested visualization, two correlation tables were generated, one is for AEVis vs. UEQ, another is for AEVis vs. UES subset (See Table 6.4).

Since we have tested two visualizations in the experiment, for all external relations, there are two cases as well. Table 6.4 summarized the results from the two cases, cells highlighted in yellow are the external correlations been investigated.

For AEVis vs. UEQ-s situation, all three relations in tested visualization no.2 have moderate to strong correlations (see upper Table 6.4(a)), and are significant at the 0.01 level (2-tailed). The correlation between hedonic pairing (AEVis hedonic aspect vs. UEQ-s hedonic quality) is 0.588; for pragmatic pairing (AEVis pragmatic aspect vs. UEQ-s pragmatic quality) is 0.690; and for total score pairing (AEVis all items vs. UEQ-s all items) is 0.747. As for tested visualization no. 3 (see upper Table 6.4(b)), two out of three relations have moderate correlations, and are significant at the 0.01 level (2-tailed). The correlation for hedonic pairing is 0.392; for pragmatic pairing is 0.544; and for total score pairing (AEVis all items vs. UEQ-s all items) is 0.646. In general, the results adequately support our expectation on the external relation between AEVis and UEQ-s; which indicates the relations between two instrument "make sense" conceptually, in light of the constructs those survey instruments were designed to assess.

As for AEVis vs. UES subset situation, both relations in tested visualization no.2 have moderate correlations (see lower Table 6.4(a)), and are significant at the 0.01 level (2-tailed). The correlation between Felt Involvement pairing (AEVis Felt Involvement related items vs. UES Felt Involvement items) is 0.672 ; for Novelty pairing (AEVis Novelty related items vs. UES Novelty items) is 0.494. As for tested visualization no. 3 (see lower Table 6.4(b)), one of the relations have moderate

Table 6.4.
Correlation tables for external relations of AEVis and other 2 related instruments:(a) correlation between AEVis and UEQ on pragmatic vs. hedonic related item sets. (b) correlation between AEVis and UES on Felt Involvement and Novelty related item sets. Cells highlighted in yellows are the external correlations we are interested in.

(a) Correlation tables for tested visualization no.2.

		AEVis		
		Hedonic	Pragmatic	Total score
UEQ	Hedonic	.588**	0.223	.486**
	Pragmatic	.375*	.690**	.546**
	Total score	.715**	.628**	.747**

		AEVis	
		FI - related	NO - related
UES	Felt Involment (FI)	.672**	.545**
	Novelty (NO)	.628**	.494**

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

(b) Correlation tables for tested visualization no.3.

		AEVis		
		Hedonic	Pragmatic	Total score
UEQ	Hedonic	.392*	.483**	.532**
	Pragmatic	.482**	.544**	.570**
	Total score	.531**	.605**	.646**

		AEVis	
		FI - related	NO - related
UES	Felt Involment (FI)	.573**	.374*
	Novelty (NO)	.587**	0.303

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

correlations, and is significant at the 0.01 level (2-tailed). The correlation between Felt Involvement pairing is 0.573 ; for Novelty pairing is only 0.303. All in all, the

results still somewhat support our expectation on the Felt Involvement-related items between AEVis and UES, which indicates the relations between the factors from two survey instruments were roughly on the same direction. However, it is not the case for novelty-related items. A potential explanation may be that the "curiosity" AEVis intend to measure is more about inquisitive or investigative behaviors, while the"curiosity" in UES is specifically limited to behaviors or actions driven by curiosity.

6.4 Findings

The external relations between AEVis and other previously established instruments (UEQ-s and UES) that are assessing related constructs were studied in Stage 4. Correlations of scores from existing instruments are able to provide additional validation support for the content of the measured construct (AE) and the generalization ability that AEVis score is likely to be replicable in the population other than the one in the field test.

In summary, the external relations been investigated are mostly positive. Correlations are stronger within the same instrument, which is expected for established surveys. While some of them are only marginally correlated, still, these results serve as a valid external validity evidence for AEVis. The construct AE is assessed by AEVis has potential to predict users' perceived usability in InfoVis and are reasonably correlated with user engagement in general.

7. DISCUSSION

7.1 Validity Evidence

Following a systematic approach of evidence-centered design [217], every activity during instrument development contributes specific evidence to support the validity of later interpretations of scores from the instrument. Followed Kane's [218] suggestion that "test scores can have multiple possible interpretations/uses, and it is the proposed interpretation/use that is validated, not the test itself or the test scores. (p. 21)" From the three stages of instrument development described in this paper, the following four types of evidences were collected to support different aspects of validity:

1. In **stage 1**, the use of grounded theory [219] to elicit user's AE characteristics from observations supports the theoretical background of the test content;
2. In **stage 2**, the feedback and review from target users and domain experts provide validity evidence for the test content of the instrument in the context of InfoVis;
3. In **stage 3**, the similar EFA structures across three different test visualizations support evidence for the internal structure of the instrument;
4. The use of IRT models confirms the results from EFA approach and further supports the score interpretation of AEVis;
5. In **stage 4**, correlations between AEVis and other established instruments were examined to support the external relation of the construct.

7.2 Using Survey Instrument AEVis

As stated in the Introduction section, AEVis is developed to be used as an evaluation tool for visualization researchers and practitioners. Especially for visualizations that are designed for communication or presentation purposes, and not for expert or high-stake serious work environment. Hence, the usage of AEVis is also oriented toward the design or development activities as described below.

The usage of AEVis is similar to NASA-TLX. NASA-TLX is a self-report, subjective assessment of user's perceived workload of a task [118]. Similarly, AEVis is also a self-report, subjective assessment of user's affective engagement of an information visualization. Both measuring constructs are self-report subjective experience with multiple facets, and the total scores are mainly to be comparative.

When using NASA-TLX, in practice, a researcher usually will have multiple treatment groups in an experiment. The researcher will compare scores from different groups (i.e. control group vs experimental group/s), then interpret the effects and interactions of the variables, or look at subscores under different sub categories. The intended use for AEVis is similar, for example, if one would like to conduct a pilot testing for a visualization prototype on a website, he or she can collect data from several participants and see their particular sub and total scores to study how they react to the visualization from the lens of AEVis. Another example, if a researcher would like to assess users' AE for a infographic, he or she might propose more than one alternatives (e.g., AB test) and try to compare the AE scores from different experiment groups, and assist visualization designers design decisions.

Inappropriate scenarios for using AEVis can be easily identified by walking though the users' tasks and the locations of the visualizations. For example, if the practitioner is creating a visualization tool for homeland security. Since the context of the workplace is high-stake and expert-oriented, the use of AEVis might not be appropriate. Or, if it an visualization artifact is hosted on a public display in a museum. Since we can expect there would be multiple users that would interact with each other in

the environment, the underlying construct of AEVis couldn't catch the consequences of multi-user interactions, it is inappropriate as well.

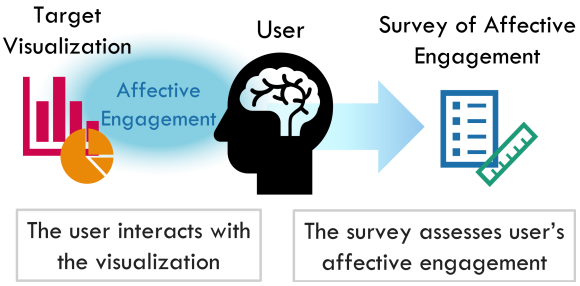
7.2.1 Use Scenario

Consider a scenario where a team of visualization practitioners want to evaluate their communicative visualization (e.g., an interactive visualization incorporated with an online magazine article) according to levels of AE within a target group of users. The practitioners can recruit a group of respondents (more is generally better, but size can be adjusted depending on resources and other factors) from their target population such as readers of an online magazine. By asking respondents to answer the items after interacting with the visualization, *the practitioner can calculate the level of AE of those particular participants. User's expected AE levels can be estimated by averaging scores for that visualization.* Figure 7.1 (a) provides a visual depiction of an evaluation scenario where a survey instrument is being used to assess AE.

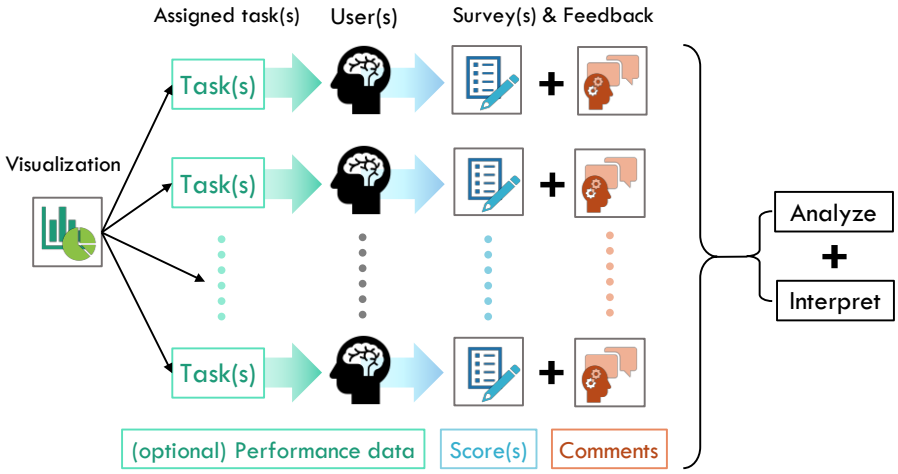
Visualization practitioners can make use of survey instruments at various stages to their design process. For instance, when a working (functional) prototype is ready, a visualization designer can conduct an evaluation session, either as a tryout or a more structured user testing described as follows:

- **Pilot tryout**: A small number of participants will be invited to use the working prototype (e.g, free exploration, tryout specific features) and then provide their comments and opinions. The items can also be used as the prompts of a short interview session. Generally, this type of quick tryout works best when the designer requires instant feedback (i.e. the work is at an early stage).
- **User Testing**: A group of participants will be recruited and asked to conduct specific tasks (e.g., solve problems, interpret visualizations, identify insights). Performance data (e.g., task accuracy and efficiency) and/or subjective responses (e.g., interview and survey) are collected via several metrics. The assigned tasks are often more structured than pilot tryout and tend to be con-

ducted when a more thorough investigation is necessary (i.e. competitor comparison).



(a) How a survey instrument assesses single user's affective engagement on one visualization.



(b) The evaluation scenario of utilizing survey instrument to assess multiple users' AE on a communicative visualization.

Figure 7.1. Use scenario of AEVis. User study results including survey instruments scores, user's subjective feedback, and (optional) user's performance data could be collected along the way.

Figure 7.1(b) is a graphical depiction of how designers can utilize a survey for a pilot tryout and for user testing. Note that for both cases, the evaluation can be conducted on-site (e.g., laboratory study) or remotely (e.g., online crowd-sourcing). A short self-report survey instrument (with roughly 10 items) will not take too much time, which makes a larger scale user testing more feasible (e.g., online crowd sourc-

ing). Even for a simple pilot tryout session, one potential benefit to employing survey instruments is that the listed items or key factors in it can stimulate rich feedback from participants. Furthermore, this scenario (see Figure 7.1(b)) also demonstrates how other performance measurements (e.g., error rate) and behavior observation methods (e.g., eye-tracking) can be integrated if more data is required.

7.2.2 Word of Caution for AEVis User

In the introduction section, I briefly elaborated why and how a self-report survey instrument that assesses AE for InfoVis can be beneficial for information visualization that focus on communicative purpose, and can be a useful option for visualization evaluation. Still, there are some limitations of this approach that should be noted in practice:

- **Interpretation:** The AEVis survey instrument is not meant to measure a visualization's AE, or any property of an artifact. Instead, it is measuring respondents' latent construct that consists of their emotional involvement as they engaged with a visualization, which is "labeled" as AE. Thus, the survey result cannot be interpreted as indicating the visualization's quality; instead, the value of the visualization against other considerations (e.g., tasks, goals) most be weighted.
- **Administration:** With different populations of respondents, the scores on the same visualization are expected to be different. Since scores here only represent *AE levels of respondents* that have been chosen, the recruitment of appropriate respondents (i.e. sampling proper participants from target population), is considerably important in influencing the results of the survey.
- **Usage:** This survey instrument is intended to measure AE for communicative and narrative visualizations, not visualizations for analysis. User intention, context, or motivation can influence AE—e.g., a high-stakes task with safety implications as part of the user's job will impact AE in ways different from

low-stakes news stories for casual users. Thus, the use of the instrument is for communicative situations or scenario where stakes are low and visualization use is focus on non-utilitarian aspects.

7.3 Concerns on AEVis as an Instrument

In the previous section, usage and word of caution of AEVis have been discussed from a practical perspective. Still, there are concerns that are more related with academic or scientific considerations. Therefore, below I layout and clarify several potential confusions reside in the gap between AE as a construct and AEVis as a survey instrument.

Dimensions of AE In stage 1, the process model of AE was illustrated as both positive and negative emotional involvement in the conceptual space. It is easy to imagine that the positive emotions such as impressed and enjoyed could result deeper emotional involvement for people. However, during the study, negative type of emotions such as worried and sad also demonstrated potential influences on participants' level of emotional involvement. Yet, considering the intended use of AEVis which is aimed to be used as an evaluation tool for assessing visualization user's level of AE. In such a scenario, potential users, including academic researchers and practitioners, are expected to be more interested on the positive type of AE. Therefore, in the follow-up study, the research team focused on the casualties between participants' activities and their emotions when participants demonstrated positive types of emotional involvement, and then elicited factors that could result AE on the positive dimension. Thus, the 11 AE indicators which eventually formed the AEVis were oriented towards the positive type of AE. That is, the AEVis instrument is designed to assess the level of "positive AE", and therefore not being able to accurately quantified the negative type of AE into scores.

Role of Performance User's performance is always an inseparable aspect for people who interact with an artifact as cognitive tasks (e.g., recognition activities, perception activities) must involve throughout the process. Therefore, the intention of developing AEVis is not to only consider the "hedonistic" aspect of human experience, nor to eliminate the entire "utilitarian" aspect of user's experience. Rather, the use of AEVis is to provided another lens to look at "user experience" considerations in the area of information visualization, with evaluation purposes in mind.

Indeed, the user performance or task efficiency should influence user's AE when they interact with a visualization. As denoted in the findings of Stage 1, the concept of AE was elicited from the casual relationships between users' various activities and emotions. That is, by definition, users' AE was constructed by both their physical (behavioral) and emotional reactions towards the target visualization. And, AEVis is an assessment tool which been designed to catch and quantified this trait into numeric scores. The mix of both hedonistic and utilitarian elements in the construct of AE is expected, and is indeed what AEVis would like to assess from the visualization users, as this is the target human latent trait to be used as the evaluation metrics.

Generalizability During the development of the AEVis, we utilized visualizations with various topics, designs, and interactive techniques in the studies. However, due to practical reasons, our tested visualizations couldn't cover all types of visualizations, nor represent the entire visualization population. The research team went through a rigorous process of instrument development (as described throughout the four stages), and tried our best to have a reasonable coverage on testbeds (tested visualizations) and proper assigned tasks during the field test. Therefore, I believe the scores calculated from AEVis have a reasonable capability to be generalized from tested participants to the general populations. still, further investigations are required to claim the generalization ability of AEVis, as generalizability across multiple populations and settings should be distinguished from generalizability to a particular or target population [220].

7.4 Deliverable of AEVis

Besides as an academic publication, there are other distribution plans for AEVis. Since the two motivations of developing AEVis are:

1. To establish a metric suitable for evaluating narrative or communicative visualizations, and
2. To provide an easy access evaluation tool for visualizations researchers and designers.

Therefore, we built a project website as an online portal of AEVis users and potential users. The website contains three types of primarily resources:

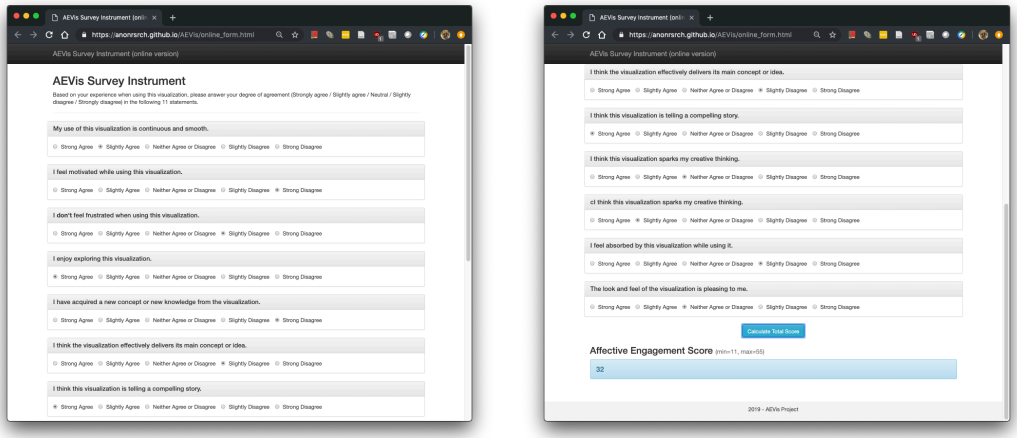
- The AEVis instrument: a paper version and a online digital version of AEVis (See Figure 7.2 and Figure 7.3)
- Supporting documents: A user manual explains the usage of AEVis and a technical report (will be released with a publication in the future)
- Analysis Tools: An Excel file that can be used to analyze AEVis responses

AEVis Survey Instrument

Based on your experience when using this visualization, please answer your degree of agreement (Strongly agree / Slightly agree / Neutral / Slightly disagree / Strongly disagree) in the following 11 statements.

Statement	Strongly Agree	Slightly Agree	Neither Agree or Disagree	Slightly Disagree	Strongly Disagree
My use of this visualization is continuous and smooth.	❑	❑	❑	❑	❑
I feel motivated while using this visualization.	❑	❑	❑	❑	❑
I **don't** feel frustrated when using this visualization.	❑	❑	❑	❑	❑
I enjoy exploring this visualization.	❑	❑	❑	❑	❑
I have acquired a new concept or new knowledge from the visualization.	❑	❑	❑	❑	❑
I think the visualization effectively delivers its main concept or idea.	❑	❑	❑	❑	❑
I think this visualization is telling a compelling story.	❑	❑	❑	❑	❑
I think this visualization sparks my creative thinking.	❑	❑	❑	❑	❑
I feel entertained when using this visualization.	❑	❑	❑	❑	❑
I feel absorbed by this visualization while using it.	❑	❑	❑	❑	❑
The look and feel of the visualization is pleasing to me.	❑	❑	❑	❑	❑

Figure 7.2. A screenshot of the paper version AEVis.



(a) The instruction and items in the online version AEVis.

(b) The items and the calculation feature of the online version AEVis.

Figure 7.3. Screenshots of online version AEVis. At the end of AEVis form, by clicking the calculation bottom, the AE score can be calculated. If the respondent misses any of the item, a reminder will be shown.

7.5 Comparisons with Existing Instruments

In recent years, there have been a number of attempts to assess engagement or related constructs in HCI domain [8]. Even though different constructs are being evaluated, it is still worth discussing the differences and similarities compared to AEVis. Below I briefly described four selected instrument that are more relevant with AEVis:

User Engagement Scale (UES) is an often cited survey instrument for general user engagement (UE) of interactive system [12]. It consists of 31 items, where each is a statement that describes the experience and the respondents have to answer their degree of agreement. UES is originally developed for online environments, thus has several items oriented towards websites or similar media. AEVis is designed specifically for InfoVis and hence has items that are specifically relevant such as visual storytelling and creativity. Moreover, the 31-item

UES instrument will take significant amount of time to answer, and therefore requires a different usage scenario than AEVis.

User Experience Questionnaire (UEQ) is a survey instrument consisting of 26 (standard) or 8 (short) items [207], and is meant to measure perceived usability (PU) of a human-computer system or interface. The shorter version (UEQ-S) has similar factor structure to AEVis, as both of them consist of hedonic and pragmatic factors in the model. By looking at the 8 items of UEQ-s, several overlapping concepts can be identified such as items measuring the exciting and interesting aspects of a system. Still, the unique context in AEVis that is particular oriented toward InfoVis in the item statement such as "effectively delivers its main concept or idea".

Engagement Scale is an engagement scale that developed for measuring user engagement in information visualization is developed by Amini [11]. There are 15 items measuring 5 different attributes of user engagement. However, due to the fact that the primary purpose of this instrument is to study and compare two different visualization types (standard vs. pictograph) and animations. The development procedure of this instrument is only a portion of the paper [11]. Thus, is considerably simplified and limited; only one small-scale field test (N = 41) has been conducted for item selection (i.e. t-test and Cronbach's alpha) before been utilized in the primary experiments. The internal structure of the measured construct and its external relations with other related metrics are not broadly examined (i.e. conducting factor analysis and correlation studies). Hence, this instrument was not investigated in the stage 4 for testing the external relations of AEVis and other related instruments.

Value of Visualization (ICE-T) Another existing instrument for evaluating visualizations is ICE-T [221]. It is a heuristic evaluation tool instead of a standard test instrument. Still, this instrument is able to generate scores for the "value" of a visualization, and to work as a grading rubric for the visualization evaluators (e.g., contest judges). Therefore, the intended users of this evaluation

tool are limited to domain experts (i.e., visualization experts). Additionally, the results interpretations are subjective to each tested visualization since some of the items only work with specific contexts such as interactions or analytical features. That is, it is possible that for visualization A, the maximum possible score will be 140; and for visualization B, the maximum possible score will be only 105. Thus, its use scenario and implications are very different from AEVis.

7.5.1 Practical Guidance for AEVis and Other Instruments

In general, most of survey or questionnaires types of measurement instruments share a similar manner or workflow of use [15]. To evaluate a visualization, the users have to recruit a group of participants, and present the target visualization to them for gathering the responses. In practice, the primary differences between AEVis and the competitors are mainly on the measured constructs (i.e. AE for AEVis, UE for UES, PU for UEQ, Quality of visualization for ICE-T). The primary consideration for selecting appropriate measurement is to identify a proper metrics per their requirement, and administrate the measurement technique accordingly [15].

Other practical differences amongst the competitors and AEVis are relatively related to the execution issues when conducting the evaluations; these issues have already been discussed in the survey of existing AE-related instruments in the literature review section. For example, the time spent for the test session might be associated with the instrument length. Longer item passages or larger number of items tend to cause longer time spent in answering the instrument (e.g., UES should take longer time than AEVis since it has more items). Another example, if the item passages contain terminologies that require domain-expertise, then the instrument users have to recruit corresponding domain experts as their respondents. That is, the intended use and target population would be changed accordingly.

Finally, if the instrument is not originally designed or developed for visualization context, modifications are needed. However, adapting an pre-existing is always

a challenge for survey or questionnaire users, it is hard to insure that the respondents will comprehend the adapted items appropriately, and therefore compromises the reliability and validity of the adapted instrument [222]. Misunderstanding of the altered items are likely to happen [223]. Moreover, by definition, the wording (e.g., [224, 225]) and the sequence of items can also contribute to the measurement construct. Altering items sequence means altering the measured constructs of a survey instrument, such item sequencing effects has been extensively discussed in the domain of test development [226, 227]. Therefore, adapting pre-existing instruments do require extra caution when using surveys or questions.

7.6 Summary

In conclusion, the rules for determining appropriate use scenario, the practical guidelines, and several scientific considerations related to the interpretation of the AEVis were discussed in this section. In practice, individual instruments are unique and has its own pros and cons. Target users, including practitioners in industry and researchers in academic, have to be cautious in the intended use of AEVis, and understand how to make proper interpretations on the calculated AEVis scores. Additionally, several interesting scientific considerations have emerged such as the dimensions of AE on its interpretation, the generalization ability of the developed survey instrument, and the relationship between the utilitarian vs. hedonian aspects within the construct of AE.

8. CONCLUSION

In this dissertation, I outlined the development and evaluation of a survey instrument for assessing Affective Engagement (AE) in InfoVis. The four main studies of our systematic process following an evidence-centered design approach were described. Plus, the use scenarios and limitations of the survey instrument were discussed and elaborated on its implications for research and practice.

- In **stage 1**, the characteristics of AE in the context of information visualization were elicited through a qualitative-driven mixed method lab study. 11 indicators for assessing AE were proposed through qualitative coding analysis.
- In **stage 2**, a quick and easy survey instrument which contain 11 items was developed based on the results of study 1. From the item bank, I went through pilot test and expert review processes to assemble the first version of our survey instrument – AEVis (Affective Engagement Visualization Survey).
- In **stage 3**, a large-scale field test with multiple rounds was conducted for instrument revisions and evaluation. Various types of analytical methods were utilized for exploring and validating the structure for the second version of AEVis.
- In **stage 4**, a follow-up field test was conducted to test external relations between AEVis and other existing instruments.
- Finally, a **discussion** section was established for the AEVis after its development. The content includes the considerations on AEVis use scenario, the work of causation for AEVis user, the competitors comparisons, and the research implications.

As InfoVis grows in popularity, non-utilitarian roles for visualization will become increasingly important. For such uses, evaluation methods based on usability and

task performance are not sufficient. Visualization researchers and designers require alternative ways to evaluate their work, especially those that assess subjective aspects of a user's experience. The instrument developed in this dissertation makes a contribution to this need. The hope is that this work can further stimulate scholarship on non-performance-related aspects of InfoVis, and that researchers and practitioners can use AEVis to assess their work quickly and easily.

For the next step, I plan to conduct a follow-up laboratory study that investigates potential correlations between AE and other factors (e.g., perceived usability, comprehension, memorability) to further study the external structure of AEVis. Also, because only a limited set of visualizations were used in the testing process, I would like to test the instrument with more types of visualizations—especially to better understand cases with lower AE scores. Additionally, from the results of EFA, the fitted model suggests that two underlying factors contribute to the target construct AE where currently were labeled as "hedonic" and "pragmatic" factors. To investigate the two factors separately, the collected data can be divided into two subsets, and analyze them with a one factor model (factor analysis) and uni-dimensional graded response model (IRT) in order to learn their internal structures as unique constructs. Moreover, for utilitarian types of visualizations, even though AE might not be a useful metric, it will be interesting to see the scores from the pragmatic-related items of AEVis.

The expectation for this research is that, by using AEVis, visualization researchers and designers can evaluate non-performance-related aspects of their work efficiently and without specific domain knowledge. Also, the utilities and implications of AE can be investigated as well. In the future, this research may expand the theoretical basis of engagement in the fields of human-computer interaction and information visualization.

REFERENCES

REFERENCES

[1] Ya-Hsin Hung and Paul Parsons. Evaluating user engagement in information visualization using mixed methods. In *IEEE VIS 17: Proceedings of the 2017 IEEE Conference on Information Visualization, Poster Abstracts*, page 2, Phoenix, AZ, 2017.

[2] Ya-Hsin Hung and Paul Parsons. Assessing User Engagement in information Visualization. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, New York, NY, USA, 2017. ACM Press.

[3] Ya-Hsin Hung and Paul Parsons. Affective engagement for communicative visualization: Quick and easy evaluation using survey instruments. In *IEEE VIS Workshop on Visualization for Communication (VisComm)*, page 6, Berlin, Germany, 2018.

[4] Rosalind W Picard and Jonathan Klein. Computers that recognise and respond to user emotion: theoretical and practical implications. *Interacting with computers*, 14(2):141–169, 2002.

[5] Marc Hassenzahl and Noam Tractinsky. User experience - a research agenda. *Behaviour & Information Technology*, 25(2):91–97, March 2006.

[6] Effie Lai-Chong Law, Virpi Roto, Marc Hassenzahl, Arnold P.O.S. Vermeeren, and Joke Kort. Understanding, scoping and defining user experience: a survey approach. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*, page 719, Boston, MA, USA, 2009. ACM Press.

[7] Donald A. Norman. *Emotional design: why we love (or hate) everyday things*. Basic Books, New York, 2005. OCLC: 254793649.

[8] Kevin Doherty and Gavin Doherty. Engagement in HCI: Conception, Theory and Measurement. *ACM Computing Surveys*, 51(5):1–39, November 2018.

[9] Narges Mahyar, Sung-Hee Kim, and Bum Chul Kwon. Towards a Taxonomy for Evaluating User Engagement in Information Visualization. In *Workshop on Personal Visualization: Exploring Everyday Life*, page 4, 2015.

[10] Bahador Saket, Carlos Scheidegger, and Stephen Kobourov. Towards Understanding Enjoyment and Flow in Information Visualization. *arXiv:1503.00582 [cs]*, March 2015. arXiv: 1503.00582.

[11] Fereshteh Amini, Nathalie Henry Riche, Bongshin Lee, Jason Leboe-McGowan, and Pourang Irani. Hooked on data videos: assessing the effect of animation and pictographs on viewer engagement. In *AVI*, pages 21–1, 2018.

[12] Heather O'Brien and Elaine G. Toms. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, 61(1):50–69, January 2010.

[13] Jeanne H. Brockmyer, Christine M. Fox, Kathleen A. Curtiss, Evan McBroom, Kimberly M. Burkhart, and Jacquelyn N. Pidruzny. The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology*, 45(4):624–634, July 2009.

[14] M. Csikszentmihalyi and R. Larson. Validity and reliability of the Experience-Sampling Method. *The Journal of Nervous and Mental Disease*, 175(9):526–536, September 1987.

[15] Mircea Fagarasanu and Shrawan Kumar. Measurement instruments and data collection: a consideration of constructs and biases in ergonomics research. *International Journal of Industrial Ergonomics*, 30(6):355 – 369, 2002.

[16] Robert J Mislevy, Russell G Almond, and Janice F Lukas. A brief introduction to evidence-centered design. *ETS Research Report Series*, 2003(1):i–29, 2003.

[17] Benjamin S Bloom et al. Taxonomy of educational objectives. vol. 1: Cognitive domain. *New York: McKay*, pages 20–24, 1956.

[18] David R Krathwohl and Lorin W Anderson. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman, 2009.

[19] *A taxonomy for learning, teaching, and assessing : a revision of Bloom's taxonomy of educational objectives /*. Longman, New York :, complete ed. edition, 2009.

[20] Fereshteh Amini, Nathalie Henry Riche, Bongshin Lee, Jason Leboe-McGowan, and Pourang Irani. Hooked on data videos: assessing the effect of animation and pictographs on viewer engagement. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces - AVI '18*, pages 1–9, Castiglione della Pescaia, Grosseto, Italy, 2018. ACM Press.

[21] Florian Windhager, Gnther Schreder, and Eva Mayr. On Inconvenient Images: Exploring the Design Space of Engaging Climate Change Visualizations for Public Audiences. *Workshop on Visualisation in Environmental Sciences (EnvirVis)*, 2019.

[22] Heather OBrien. Theoretical Perspectives on User Engagement. In Heather O'Brien and Paul Cairns, editors, *Why Engagement Matters*, pages 1–26. Springer International Publishing, Cham, 2016.

[23] Bahador Saket, Alex Endert, and John Stasko. Beyond Usability and Performance: A Review of User Experience-focused Evaluations in Visualization. In *Proceedings of the Beyond Time and Errors on Novel Evaluation Methods for Visualization - BELIV '16*, pages 133–142, New York, New York, USA, 2016. ACM Press.

[24] Jeremy Boy, Francoise Detienne, and Jean-Daniel Fekete. Storytelling in Information Visualizations: Does It Engage Users to Explore Data? In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1449–1458, New York, NY, USA, 2015. ACM.

[25] Andrew Vande Moere, Martin Tomitsch, Christoph Wimmer, Boesch Christoph, and Thomas Grechenig. Evaluating the Effect of Style in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2739–2748, December 2012.

[26] Heather O'Brien and Elaine G. Toms. What is User Engagement? A Conceptual Framework for Defining User Engagement with Technology. *Journal of the American Society for Information Science and Technology*, 59(6):938–955, April 2008.

[27] Alistair Sutcliffe. Designing for User Engagement: Aesthetic and Attractive User Interfaces. *Synthesis Lectures on Human-Centered Informatics*, 2(1):1–55, January 2009.

[28] Phil Turner. The Anatomy of Engagement. In *Proceedings of the 28th Annual European Conference on Cognitive Ergonomics*, ECCE '10, pages 59–66, New York, NY, USA, 2010. ACM.

[29] Simon Attfield, Gabriella Kazai, Mounia Lalmas, and Benjamin Piwowarski. Towards a science of user engagement (position paper). In *WSDM Workshop on User Modelling for Web Applications*, 2011.

[30] Robert Kosara. Presentation-oriented visualization techniques. *IEEE computer graphics and applications*, 36(1):80–85, 2016.

[31] D. Betsy McCoach, Robert K. Gable, and John P. Madura. *Instrument Development in the Affective Domain: School and Corporate Applications*. Springer Science & Business Media, May 2013.

[32] Martin EP Seligman and Mihaly Csikszentmihalyi. *Positive psychology: An introduction*. Springer, 2014.

[33] Mihaly Csikszentmihalyi, Sami Abuhamdeh, and Jeanne Nakamura. Flow. In *Flow and the Foundations of Positive Psychology*, pages 227–238. Springer Netherlands, 2014.

[34] Andrew J Martin and Martin Dowson. Interpersonal relationships, motivation, engagement, and achievement: Yields for theory, current issues, and educational practice. *Review of educational research*, 79(1):327–365, 2009.

[35] Heather Davis, Jessica Summers, and Lauren Miller. *An Interpersonal Approach to Classroom Management: Strategies for Improving Student Engagement*. Corwin Press, 2590 Conejo Spectrum, Thousand Oaks California 91320 United States, 2012.

[36] Janette Lehmann, Mounia Lalmas, Elad Yom-Tov, and Georges Dupret. Models of User Engagement. In *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization*, UMAP'12, pages 164–175, Berlin, Heidelberg, 2012. Springer-Verlag.

[37] Kerry Rodden, Hilary Hutchinson, and Xin Fu. Measuring the User Experience on a Large Scale: User-centered Metrics for Web Applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 2395–2398, New York, NY, USA, 2010. ACM.

[38] Yellowlees Douglas and Andrew Hargadon. The Pleasure Principle: Immersion, Engagement, Flow. In *Proceedings of the Eleventh ACM on Hypertext and Hypermedia*, HYPERTEXT '00, pages 153–160, New York, NY, USA, 2000. ACM.

[39] Mihaly Csikszentmihalyi. The flow experience and its significance for human psychology. In *Optimal experience*. Cambridge University Press, 1988.

[40] Emily Brown and Paul Cairns. A Grounded Investigation of Game Immersion. In *Proceedings of ACM CHI 2004 Conference on Human Factors in Computing Systems*, pages 1297–1300. Press, 2004.

[41] Penelope Sweetser and Peta Wyeth. GameFlow: a model for evaluating player enjoyment in games. *Computers in Entertainment*, 3(3):3, July 2005.

[42] Andrës Lucero, Jussi Holopainen, Elina Ollila, Riku Suomela, and Evangelos Karapanos. The Playful Experiences (PLEX) Framework As a Guide for Expert Evaluation. In *Proceedings of the 6th International Conference on Designing Pleasurable Products and Interfaces*, DPPI '13, pages 221–230, New York, NY, USA, 2013. ACM.

[43] Juha Arrasvuori, Hannu Korhonen, and Kaisa Väänänen-Vainio-Mattila. Exploring Playfulness in User Experience of Personal Mobile Products. In *Proceedings of the 22Nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction*, OZCHI '10, pages 88–95, New York, NY, USA, 2010. ACM. 00012.

[44] Juha Arrasvuori, Marion Boberg, Jussi Holopainen, Hannu Korhonen, Andrës Lucero, and Markus Montola. Applying the PLEX Framework in Designing for Playfulness. In *Proceedings of the 2011 Conference on Designing Pleasurable Products and Interfaces*, DPPI '11, pages 24:1–24:8, New York, NY, USA, 2011. ACM. 00029.

[45] Heather O'Brien, Luanne Freund, and Richard Kopak. Investigating the Role of User Engagement in Digital Reading Environments. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, pages 71–80, New York, NY, USA, 2016. ACM.

[46] Gregory J Boyle, Edward Helmes, Gerald Matthews, and Carroll E Izard. Measures of affect dimensions. In *Measures of personality and social psychological constructs*, pages 190–224. Elsevier, 2015.

[47] Paul Ekman. Moods, emotions, and traits. In P. Eckman and R. Davidson, editors, *The Nature of Emotion*. 1994.

[48] Cynthia D Fisher. Mood and emotions while working: missing pieces of job satisfaction? *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, 21(2):185–202, 2000.

[49] Liam J. Bannon. From Human Factors to Human Actors: The Role of Psychology and Human-Computer Interaction Studies in System Design. In RONALD M. BAECKER, JONATHAN GRUDIN, WILLIAM A. S. BUXTON, and SAUL GREENBERG, editors, *Readings in HumanComputer Interaction*, Interactive Technologies, pages 205 – 214. Morgan Kaufmann, 1995.

[50] Susanne Bdker. When second wave HCI meets third wave challenges. In *Proceedings of the 4th Nordic conference on Human-computer interaction changing roles - NordiCHI '06*, pages 1–8, Oslo, Norway, 2006. ACM Press.

[51] Susanne Bdker. Third-wave HCI, 10 years later-participation and sharing. pages 24–31, September 2015.

[52] Donald A. Norman. Emotion & Design: Attractive things work better, November 2008.

[53] James F. Sallis and Brian E. Saelens. Assessment of Physical Activity by Self-Report: Status, Limitations, and Future Directions. *Research Quarterly for Exercise and Sport*, 71(sup2):1–14, June 2000.

[54] Louisa G. Sylvia, Emily E. Bernstein, Jane L. Hubbard, Leigh Keating, and Ellen J. Anderson. Practical Guide to Measuring Physical Activity. *Journal of the Academy of Nutrition and Dietetics*, 114(2):199–208, February 2014.

[55] Delroy L. Paulhus and Simine Vazire. The self-report method. *Handbook of research methods in personality psychology*, 1:224–239, 2007.

[56] Robert J. Fisher. Social Desirability Bias and the Validity of Indirect Questioning. *Journal of Consumer Research*, 20(2):303, September 1993.

[57] Maryon F. King and Gordon C. Bruner. Social desirability bias: A neglected aspect of validity testing. *Psychology and Marketing*, 17(2):79–103, February 2000.

[58] Roger E. Millsap and Howard T. Everson. Methodology Review: Statistical Approaches for Assessing Measurement Bias. *Applied Psychological Measurement*, 17(4):297–334, December 1993.

[59] Mounia Lalmas, Heather O'Brien, and Elad Yom-Tov. Measuring User Engagement. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 6(4):1–132, 2014.

[60] Norman K. Denzin and Yvonna S. Lincoln, editors. *Collecting and interpreting qualitative materials*. Sage Publications, Thousand Oaks, Calif, 3rd ed edition, 2008.

[61] Barbara DiCicco-Bloom and Benjamin F Crabtree. The qualitative research interview. *Medical Education*, 40(4):314–321, April 2006.

[62] Heather L OBrien. Measuring user engagement with information systems. *Dalhousie University, Canada*, 2006.

[63] Jonathan A. Tran, Katie S. Yang, Katie Davis, and Alexis Hiniker. Modeling the engagement-disengagement cycle of compulsive phone use. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 312:1–312:14, New York, NY, USA, 2019. ACM.

[64] Melanie Kellar, Carolyn Watters, and Kori M. Inkpen. An exploration of web-based monitoring: Implications for design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 377–386, New York, NY, USA, 2007. ACM.

[65] Sinh Huynh, Seungmin Kim, JeongGil Ko, Rajesh Krishna Balan, and Youngki Lee. Engagemon: Multi-modal engagement sensing for mobile games. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(1):13:1–13:27, March 2018.

[66] Saskia M. Kelders and Hanneke Kip. Development and initial validation of a scale to measure engagement with ehealth technologies. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, pages LBW0185:1–LBW0185:6, New York, NY, USA, 2019. ACM.

[67] Jakob Nielsen. Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human-Computer Studies*, 41(3):385–397, September 1994.

[68] Erica L. Olmsted-Hawala, Elizabeth D. Murphy, Sam Hawala, and Kathleen T. Ashenfelter. Think-aloud protocols: a comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. page 2381. ACM Press, 2010.

[69] K. Anders Ericsson and Herbert A. Simon. Verbal reports as data. *Psychological Review*, 87(3):215–251, 1980.

[70] K Anders Ericsson. *Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts performance on representative tasks*. 2006.

[71] Maaike van den Haak, Menno De Jong, and Peter Jan Schellens. Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology*, 22(5):339–351, September 2003.

[72] T. Boren and J. Ramey. Thinking aloud: reconciling theory and practice. *IEEE Transactions on Professional Communication*, 43(3):261–278, September 2000.

[73] Janni Nielsen, Torkil Clemmensen, and Carsten Yssing. Getting access to what goes on in people's heads?: reflections on the think-aloud technique. page 101. ACM Press, 2002.

[74] Sarah Leon Rojas, Leif Oppermann, Lisa Blum, and Martin Wolpers. Natural europe educational games suite: Using structured museum-data for creating mobile educational games. In *Proceedings of the 11th Conference on Advances in Computer Entertainment Technology*, ACE '14, pages 6:1–6:6, New York, NY, USA, 2014. ACM.

[75] Chek Tien Tan, Tuck Wah Leong, Songjia Shen, Christopher Dubravs, and Chen Si. Exploring gameplay experiences on the oculus rift. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '15, pages 253–263, New York, NY, USA, 2015. ACM.

[76] Charlene Jennett, Anna L. Cox, Paul Cairns, Samira Dhoparee, Andrew Epps, Tim Tijs, and Alison Walton. Measuring and Defining the Experience of Immersion in Games. *Int. J. Hum.-Comput. Stud.*, 66(9):641–661, September 2008.

[77] Charlene Jennett, Anna L. Cox, and Paul Cairns. Investigating Computer Game Immersion and the Component Real World Dissociation. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '09, pages 3407–3412, New York, NY, USA, 2009. ACM.

[78] Giovanni B. Moneta. On the Measurement and Conceptualization of Flow. In Stefan Engeser, editor, *Advances in Flow Research*, pages 23–50. Springer New York, 2012.

[79] Jane Lessiter, Jonathan Freeman, Edmund Keogh, and Jules Davidoff. A Cross-Media Presence Questionnaire: The ITC-Sense of Presence Inventory. *Presence: Teleoperators and Virtual Environments*, 10(3):282–297, June 2001.

[80] Andrës Lucero and Juha Arrasvuori. PLEX Cards: A Source of Inspiration when Designing for Playfulness. In *Proceedings of the 3rd International Conference on Fun and Games*, Fun and Games '10, pages 28–37, New York, NY, USA, 2010. ACM.

[81] Ioannis Arapakis, Konstantinos Athanasakos, and Joemon M. Jose. A comparison of general vs personalised affective models for the prediction of topical relevance. page 371. ACM Press, 2010.

[82] Serina A Neumann and Shari R Waldstein. Similar patterns of cardiovascular response during emotional activation as a function of affective valence and arousal and gender. *Journal of Psychosomatic Research*, 50(5):245–253, May 2001.

[83] Ichiro Uchiyama. Differentiation of fear, anger, and joy. *Perceptual and Motor Skills*, 74(2):663–667, 1992.

[84] Sylvia D. Kreibig. Autonomic nervous system activity in emotion: A review. *Biological Psychology*, 84(3):394–421, July 2010.

[85] Jerry S. Wiggins. *Personality and prediction: Principles of personality assessment*. Addison-Wesley, 1973.

[86] C. Ghaoui. eye-tracking in HCI and usability Research. In *Encyclopedia of Human Computer Interaction*, pages 211–219. Idea Group Reference, 2005.

[87] Jeff Klingner, Rakshit Kumar, and Pat Hanrahan. Measuring the task-evoked pupillary response with a remote eye tracker. page 69. ACM Press, 2008.

[88] TobiiPro.com Learning Center. Dark and bright pupil tracking, August 2015.

[89] Martin H. Fischer. An Investigation of Attention Allocation During Sequential Eye Movement Tasks. *The Quarterly Journal of Experimental Psychology Section A*, 52(3):649–677, August 1999.

[90] James E. Hoffman and Baskaran Subramaniam. The role of visual attention in saccadic eye movements. *Perception & Psychophysics*, 57(6):787–795, January 1995.

[91] Heiner Deubel and Werner X. Schneider. Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, 36(12):1827–1837, June 1996.

[92] Michael D. Byrne, John R. Anderson, Scott Douglass, and Michael Matessa. Eye tracking the visual search of click-down menus. pages 402–409. ACM Press, 1999.

[93] Georg Buscher, Susan T. Dumais, and Edward Cutrell. The good, the bad, and the random: an eye-tracking study of ad quality in web search. page 42. ACM Press, 2010.

[94] Ioannis Arapakis, Mounia Lalmas, B. Barla Cambazoglu, Mari-Carmen Marcos, and Joemon M. Jose. User engagement in online News: Under the scope of sentiment, interest, affect, and gaze: User Engagement in Online News: Under the Scope of Sentiment, Interest, Affect, and Gaze. *Journal of the Association for Information Science and Technology*, 65(10):1988–2005, October 2014.

[95] Daniel Kahneman, Bernard Tursky, David Shapiro, and Andrew Crider. Pupillary, heart rate, and skin resistance changes during a mental task. *Journal of Experimental Psychology*, 79(1, Pt.1):164–167, 1969.

[96] S.P. Marshall. The Index of Cognitive Activity: measuring cognitive workload. pages 7–5–7–9. IEEE, 2002.

[97] Shamsi T. Iqbal, Piotr D. Adamczyk, Xianjun Sam Zheng, and Brian P. Bailey. Towards an index of opportunity: understanding changes in mental workload during task execution. page 311. ACM Press, 2005.

[98] D. Kahneman and J. Beatty. Pupil Diameter and Load on Memory. *Science*, 154(3756):1583–1585, December 1966.

[99] Jackson Beatty and Daniel Kahneman. Pupillary changes in two memory tasks. *Psychonomic Science*, 5(10):371–372, October 1966.

[100] Margaret M. Bradley, Laura Miccoli, Miguel A. Escrig, and Peter J. Lang. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45(4):602–607, July 2008.

[101] Laura A. Granka, Thorsten Joachims, and Geri Gay. Eye-tracking analysis of user behavior in WWW search. page 478. ACM Press, 2004.

[102] Bing Pan, Helene A. Hembrooke, Geri K. Gay, Laura A. Granka, Matthew K. Feusner, and Jill K. Newman. The determinants of web page viewing behavior: an eye-tracking study. pages 147–154. ACM Press, 2004.

[103] Jeff Huang, Ryen W. White, and Susan Dumais. No clicks, no problem: using cursor movements to understand and improve search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1225–1234. ACM, 2011.

[104] Mon Chu Chen, John R. Anderson, and Myeong Ho Sohn. What can a mouse cursor tell us more?: correlation of eye/mouse movements on web browsing. page 281. ACM Press, 2001.

[105] Qi Guo and Eugene Agichtein. Towards predicting web searcher gaze position from mouse movements. page 3601. ACM Press, 2010.

[106] Edward Cutrell and Zhiwei Guan. What are you looking for?: an eye-tracking study of information usage in web search. page 407. ACM Press, 2007.

[107] Qi Guo and Eugene Agichtein. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. page 569. ACM Press, 2012.

[108] Georges Dupret and Mounia Lalmas. Absence time and user engagement: evaluating ranking functions. page 173. ACM Press, 2013.

[109] Heather O'Brien and Paul Cairns. An empirical evaluation of the User Engagement Scale (UES) in online news environments. *Information Processing & Management*, 51(4):413–427, 2015.

[110] Paul Thomas, Heather O'Brien, and Tom Rowlands. Measuring Engagement with Online Forms. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, pages 325–328, New York, NY, USA, 2016. ACM.

[111] Thomas Tullis and William Albert. *Measuring the User Experience, Second Edition: Collecting, Analyzing, and Presenting Usability Metrics*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2nd edition, 2013.

[112] Jeff Sauro and James R. Lewis. *Quantifying the user experience : practical statistics for user research*. Amsterdam, Netherlands : Morgan Kaufmann, second edi edition, 2016.

[113] Ben Shneiderman, Catherine Plaisant, Maxine Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. *Designing the user interface: strategies for effective human-computer interaction*. Pearson, 2016.

[114] J. P. Chin, V. A. Diehl, and L. K. Norman. Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '88*, pages 213–218, New York, New York, USA, 1988. ACM Press.

[115] Arnold Lund. *Measuring Usability with the USE Questionnaire*, volume 8. jan 2001.

[116] Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, 27(3):425, 2003.

[117] John Brooke. SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[118] Sandra G Hart and Lowell E Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In Peter A Hancock and Najmedin B T Advances in Psychology Meshkati, editors, *Human Mental Workload*, volume 52, pages 139–183. North-Holland, 1988.

[119] Fred D Davis. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, 13(3):319–340, 1989.

[120] James R. Lewis and James R. Psychometric evaluation of an after-scenario questionnaire for computer usability studies. *ACM SIGCHI Bulletin*, 23(1):78–81, dec 1990.

[121] James R. Lewis. Psychometric Evaluation of the Post-Study System Usability Questionnaire: The PSSUQ. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 36(16):1259–1260, oct 1992.

[122] Jakob Nielsen. Usability inspection methods. pages 413–414. ACM, 1994.

[123] James R. Lewis and James R. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7(1):57–78, jan 1995.

[124] Jurek Kirakowski and Mary Corbett. *SUMI: the Software Usability Measurement Inventory*, volume 24. oct 2006.

[125] Bob G. Witmer and Michael J. Singer. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoperators & Virtual Environments*, 7(3):225–240, June 1998.

[126] Jurek Kirakowski and Bozena Cierlik. Measuring the Usability of Web Sites. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 42(4):424–428, oct 1998.

[127] Afke Donker. Human factors in educational software for young children. page 136, 2005.

[128] Raafat Saadé and Bouchaib Bahli. The impact of cognitive absorption on perceived usefulness and perceived ease of use in on-line learning: an extension of the technology acceptance model. *Information & Management*, 42(2):317–327, 2005.

[129] Donna P. Tedesco and Thomas S. Tullis. A Comparison of Methods for Eliciting Post-Task Subjective Ratings in Usability Testing. 2006.

[130] Bettina Laugwitz, Theo Held, and Martin Schrepp. Construction and Evaluation of a User Experience Questionnaire. pages 63–76. Springer, Berlin, Heidelberg, nov 2008.

[131] Jeff Sauro and Joseph S. Dumas. Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*, page 1599, New York, New York, USA, 2009. ACM Press.

[132] Jeeyun Oh, Saraswathi Bellur, and S Shyam Sundar. Clicking, Assessing, Immersing, and Sharing: An Empirical Model of User Engagement with Interactive Media. *Communication Research*, 45(5):737–763, sep 2015.

[133] Jeff Sauro. SUPR-Q: A Comprehensive Measure of the Quality of the Website User Experience. *J. Usability Studies*, 10(2):68–86, feb 2015.

[134] Heather O'Brien. Theoretical perspectives on user engagement. In *Why Engagement Matters*, pages 1–26. Springer International Publishing, 2016.

[135] Bahador Saket, Alex Endert, and John Stasko. Beyond usability and performance: A review of user experience-focused evaluations in visualization. In *Proceedings of the Beyond Time and Errors on Novel Evaluation Methods for Visualization - BELIV '16*, pages 133–142, New York, USA, 2016. ACM Press.

[136] Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE transactions on visualization and computer graphics*, 18(9):1520–1536, 2012.

[137] Kelly Caine. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, pages 981–992, Santa Clara, California, USA, 2016. ACM Press.

[138] Juliet M. Corbin and Anselm Strauss. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, 13(1):3–21, 1990.

[139] Nancy Carter, Denise Bryant-Lukosius, Alba DiCenso, Jennifer Blythe, and Alan J. Neville. The Use of Triangulation in Qualitative Research. *Oncology Nursing Forum*, 41(5):545–547, September 2014.

[140] Michael Quinn Patton. Enhancing the quality and credibility of qualitative analysis. *Health services research*, 34(5 Pt 2):1189, 1999.

[141] Sharlene Hesse-Biber. Emerging methodologies and methods practices in the field of mixed methods research. *Qualitative Inquiry*, 16:415–418, 2010.

[142] Tobii AB. Tobii Studio, 2017.

[143] Tobii AB. Working with Heat Maps and Gaze Plots, November 2015.

[144] Christopher J Johnstone, Nicole A Bottsford-Miller, and Sandra J Thompson. Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and english language learners. technical report 44. *National Center on Educational Outcomes, University of Minnesota*, 2006.

[145] Jennifer L Branch. Investigating the information-seeking processes of adolescents: The value of using think alouds and think afters. *Library & Information Science Research*, 22(4):371–392, November 2000.

[146] K. Anders Ericsson and Herbert A. Simon. How to Study Thinking in Everyday Life: Contrasting Think-Aloud Protocols With Descriptions and Explanations of Thinking. *Mind, Culture, and Activity*, 5(3):178–186, July 1998.

[147] Keshif, LLC. U.S. Presidents, 2014.

[148] Adam McCann. Oscar Predictions, February 2014.

[149] ErrantScience.com. When to eat chocolate: A guide for Researchers, 2016.

[150] Nathan Yau. Causes of death, January 2016.

[151] Ellen F. Olshansky. Theoretical Issues in Building a Grounded Theory: Application of an Example of a Program of Research on Infertility. *Qualitative Health Research*, 6(3):394–405, August 1996.

[152] Johnny Saldana. *The coding manual for qualitative researchers.* SAGE, Los Angeles, [Calif.] ; London, 2009.

[153] Alonso H. Vera and Herbert A. Simon. Situated action: A symbolic interpretation. *Cognitive Science*, 17(1):7–48, January 1993.

[154] Robert L. Ebel. *Essentials of educational measurement.* Essentials of educational measurement. Prentice-Hall, Oxford, England, 1972.

[155] Robert M. Thorndike, George K. Cunningham, Robert Ladd Thorndike, and Elizabeth P. Hagen. *Measurement and evaluation in psychology and education, 5th ed.* Measurement and evaluation in psychology and education, 5th ed. Macmillan Publishing Co, Inc, New York, NY, England, 1991.

[156] Robert F DeVellis. *Scale development: Theory and applications*, volume 26. Sage publications, 2016.

[157] Robert M. Thorndike and Tracy M. Thorndike-Christ. *Measurement and Evaluation in Psychology and Education*. Pearson, Boston, 8 edition edition, May 2009.

[158] Ronald A. Berk. Importance of expert judgment in content-related validity evidence. *Western Journal of Nursing Research*, 12(5):659–671, 1990. PMID: 2238643.

[159] R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 22 140:55–55, 1932.

[160] Deborah L. Bandalos. *Measurement theory and applications for the social sciences*. Methodology in the social sciences. The Guilford Press, 2018.

[161] Peter G Polson, Clayton Lewis, John Rieman, and Cathleen Wharton. Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. *International Journal of man-machine studies*, 36(5):741–773, 1992.

[162] David A Harrison, Mary E McLaughlin, and Terry M Coalter. Context, cognition, and common method variance: Psychometric and verbal protocol evidence. *Organizational Behavior and Human Decision Processes*, 68(3):246–261, 1996.

[163] M. R. Lynn. Determination and quantification of content validity. *Nursing Research*, 35(6):382–385, December 1986.

[164] Qualtrics. Qualtrics XM, 2005.

[165] Joan S. Grant and Linda L. Davis. Selection and use of content experts for instrument development. *Research in Nursing & Health*, 20(3):269–274, June 1997.

[166] Amazon Mechanical Turk. https://www.mturk.com.

[167] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. Running experiments on amazon mechanical turk. 2010.

[168] Kyle A. Thomas and Scott Clifford. Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, 77:184–197, 2017.

[169] RJ Andrews. Creative routines [infographic], 2014. Retrieved March 15, 2019 from https://www.informationisbeautifulawards.com/showcase/483-creative-routines.

[170] Matthew Weber and Maryanne Murray. Fifa world cup tournaments, 2018. Retrieved March 15, 2019 from http://fingfx.thomsonreuters.com/gfx/rngs/SPORTS-WORLDCUP/010051ZL4GR/index.html.

[171] Arthur Soares, Beatrys Rodrigues, Cintia Ferreira, Laura Del Vecchio, Lidia Zuin, Lucas Munhoz, and Thomaz Rezende. 2001: A space odyssey, 2017. Retrieved March 15, 2019 fromhttps://viz.envisioning.io/2001aspaceodyssey/?o=0.

[172] Sandra Ferketich. Focus on psychometrics. aspects of item analysis. *Research in nursing & health*, 14(2):165–168, 1991.

[173] J.C. Nunnally. *Psychometric theory*. McGraw-Hill series in psychology. McGraw-Hill, 2nd edition, 1978.

[174] Linda Lindsey Davis. Instrument review: Getting the most from a panel of experts. *Applied Nursing Research*, 5(4):194–197, November 1992.

[175] Barbara S. Plake. A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process. *Educational and Psychological Measurement*, 40(2):397–404, 1980.

[176] Ira H Bernstein and Jum C Nunnally. Psychometric theory. *New York: McGraw-Hill. Oliva, TA, Oliver, RL, & MacMillan, IC (1992). A catastrophe model for developing service satisfaction strategies. Journal of Marketing*, 56:83–95, 1994.

[177] M.A. Pett, N.R. Lackey, and J.J. Sullivan. *Making Sense of Factor Analysis: The Use of Factor Analysis for Instrument Development in Health Care Research*. SAGE Publications, 2003.

[178] Richard L. Gorsuch. Exploratory Factor Analysis. In John R. Nesselroade and Raymond B. Cattell, editors, *Handbook of Multivariate Experimental Psychology*, pages 231–258. Springer US, Boston, MA, 1988.

[179] Robin K. Henson and J. Kyle Roberts. Use of Exploratory Factor Analysis in Published Research: Common Errors and Some Comment on Improved Practice. *Educational and Psychological Measurement*, 66(3):393–416, June 2006.

[180] Richard L. Gorsuch. Exploratory Factor Analysis: Its Role in Item Analysis. *Journal of Personality Assessment*, 68(3):532–560, June 1997.

[181] Barbara G Tabachnick, Linda S Fidell, and Jodie B Ullman. *Using multivariate statistics*, volume 5. Pearson Boston, MA, 2007.

[182] J. C. F. de Winter, D. Dodou*, and P. A. Wieringa. Exploratory Factor Analysis With Small Sample Sizes. *Multivariate Behavioral Research*, 44(2):147–181, April 2009.

[183] Andrew L. Comrey. *A First Course in Factor Analysis*. Psychology Press, 2 edition, November 2013.

[184] Larry G Daniel. Comparisons of exploratory and confirmatory factor analysis. 1989.

[185] David W. Gerbing and Janet G. Hamilton. Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 3(1):62–72, January 1996.

[186] Nambury S. Roju, Wim J. van der Linden, and Paul F. Fleer. Irt-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, 19(4):353–368, 1995.

[187] Pere J. Ferrando and Eliseo Chico. The construct of sensation seeking as measured by Zuckerman's SSS-V and Arnett's AISS: a structural equation model. *Personality and Individual Differences*, 31(7):1121–1133, November 2001.

[188] Shengyu Jiang, Chun Wang, and David J. Weiss. Sample Size Requirements for Estimation of Item Parameters in the Multidimensional Graded Response Model. *Frontiers in Psychology*, 7:109, 2016.

[189] Maria Orlando Edelen and Bryce B. Reeve. Applying item response theory (irt) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, 16(1):5, Mar 2007.

[190] Cheryl T. Beck and Robert K. Gable. Item Response Theory in Affective Instrument Development: An Illustration. *Journal of Nursing Measurement*, 9(1):5–22, May 2001.

[191] William Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois, 2018. R package version 1.8.12.

[192] John Fox. *polycor: Polychoric and Polyserial Correlations*, 2016. R package version 0.7-9.

[193] Tony CM Lam and Joseph J Stevens. Effects of content polarization, item wording, and rating scale width on rating response. *Applied Measurement in Education*, 7(2):141–158, 1994.

[194] Gail H. Weems, Anthony J. Onwuegbuzie, and Kathleen M.T. Collins. The Role of Reading Comprehension in Responses to Positively and Negatively Worded Items on Rating Scales. *Evaluation & Research in Education*, 19(1):3–20, February 2006.

[195] Jason C Chan. Response-order effects in likert-type scales. *Educational and Psychological Measurement*, 51(3):531–540, 1991.

[196] An Gie Yong and Sean Pearce. A beginners guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in quantitative methods for psychology*, 9(2):79–94, 2013.

[197] Robin K Henson and J Kyle Roberts. Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological measurement*, 66(3):393–416, 2006.

[198] Leandre R Fabrigar, Duane T Wegener, Robert C MacCallum, and Erin J Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, 4(3):272, 1999.

[199] Litze Hu and Peter M. Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1):1–55, January 1999.

[200] Roderick P. McDonald. *Test theory: A unified treatment.* Test theory: A unified treatment. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 1999.

[201] William Revelle and Richard E Zinbarg. Coefficients alpha, beta, omega, and the glb: Comments on sijtsma. *Psychometrika*, 74(1):145, 2009.

[202] Lifang Deng and Wai Chan. Testing the Difference Between Reliability Coefficients Alpha and Omega. *Educational and Psychological Measurement*, 77(2):185–203, April 2017.

[203] Lee J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, September 1951.

[204] Thomas J. Dunn, Thom Baguley, and Vivienne Brunsden. From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, 105(3):399–412, August 2014.

[205] Jason W Osborne, Anna B Costello, and J Thomas Kellow. Best practices in exploratory factor analysis. *Best practices in quantitative methods*, pages 86–99, 2008.

[206] Coen A. Bernaards and Robert I.Jennrich. Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement*, 65:676–696, 2005.

[207] Bettina Laugwitz, Theo Held, and Martin Schrepp. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and Usability Engineering Group*, pages 63–76. Springer, 2008.

[208] T.A. Brown. *Confirmatory Factor Analysis for Applied Research, Second Edition.* Methodology in the Social Sciences. Guilford Publications, 2014.

[209] James B. Schreiber, Amaury Nora, Frances K. Stage, Elizabeth A. Barlow, and Jamie King. Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. *The Journal of Educational Research*, 99(6):323–338, July 2006.

[210] Yves Rosseel. lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36, 2012.

[211] Sacha Epskamp. *semPlot: Path Diagrams and Visual Analysis of Various SEM Packages' Output*, 2019. R package version 1.1.1.

[212] R J de Ayala. *The Theory and Practice of Item Response Theory.* Methodology in the Social Sciences. Guilford Publications, 2013.

[213] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, December 1974.

[214] Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, March 1978.

[215] Li Cai, David Thissen, and Stephen du Toit. IRTPRO 4.20 [Computer software]. Retrieved March 31, 2019 from http://www.ssicentral.com/index.php/products/irt.

[216] IBM Corp. IBM SPSS Statistics for Windows, 2017.

[217] Robert J Mislevy and Geneva D Haertel. Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, 25(4):6–20, 2006.

[218] Michael T Kane. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1):1–73, 2013.

[219] Paul Cairns and Anna L. Cox. A qualititative approach to HCI research. In *Research methods for human-computer interaction*, pages 138–157. Cambridge University Press, Cambridge, UK ; New York, 2008. OCLC: ocn212858871.

[220] H.T. Reis and C.M. Judd. *Handbook of Research Methods in Social and Personality Psychology*. Handbook of Research Methods in Social and Personality Psychology. Cambridge University Press, 2000.

[221] Emily Wall, Meeshu Agnihotri, Laura Matzen, Kristin Divis, Michael Haass, Alex Endert, and John Stasko. A heuristic approach to value-driven evaluation of visualizations. *IEEE transactions on visualization and computer graphics*, 25(1):491–500, 2019.

[222] Robert J. McDermott and Paul D. Sarvela. *Health education evaluation and measurement: a practitioner's perspective*. WCB/McGraw-Hill, Dubuque, IA, 2nd ed edition, 1999.

[223] Vanessa E. C. Sousa, Jeffrey Matson, and Karen Dunn Lopez. Questionnaire Adapting: Little Changes Mean a Lot. *Western Journal of Nursing Research*, 39(9):1289–1300, September 2017.

[224] Howard Schuman and Stanley Presser. Question Wording as an Independent Variable in Survey Analysis. *Sociological Methods & Research*, 6(2):151–170, November 1977.

[225] Chester A. Schriesheim and Regina J. Eisenbach. An Exploratory and Confirmatory Factor-Analytic Investigation of Item Wording Effects on the Obtained Factor Structures of Survey Questionnaire Measures. *Journal of Management*, 21(6):1177–1193, December 1995.

[226] Linda F. Leary and Neil J. Dorans. Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research*, 55(3):387–413, 1985.

[227] Meghan Rector Federer, Ross H. Nehm, John E. Opfer, and Dennis Pearl. Using a Constructed-Response Instrument to Explore the Effects of Item Position and Item Features on the Assessment of Students Written Scientific Explanations. *Research in Science Education*, 45(4):527–553, August 2015.

[228] Peter Gilks. Record Breaking Coasters, November 2016.

[229] Rody Zakovich. Analysis of Queen, December 2016.

[230] Brit Cava. Women in Tech: Who's Receiving Computer Science Degrees?, November 2016.

[231] Mike Bostock. Hierarchical edge bundling, 2011.

[232] DataAddict. From 1945 to 2015: 70 years of first names in France.

[233] Gapminder.com. Gapminder Tools, May 2017.

[234] Takashi Ohno. Wealth Inequality in the US, November 2016.

[235] Ella Koeze. 35 Years Of American Death, dec 2017.

[236] TW Gonzalez. 2015 NFL Predictions, November 2015.

[237] LLC KESHIF. Medals Won by Olympic Athletes, 2017.

[238] Curtis Harris. New York Taxis, November 2016.

[239] NeoMam Studios. The Evolution of the Office Desk, 2016.

[240] Bernette Becker. Sugar: The Bitter Truth, 2015.

[241] Christian Tate. Avengers Assemble, 2013.

[242] Ian H. Yoo. Say it in Engrish!, 2013.

[243] CustomMade. Carbon Footprint of the Internet, April 2015.

[244] Madison Rosa. 10 User Engagement Metrics Everyone Can Use, 2014.

[245] Fabio Bergamaschi. Hard Knuckles - Top 500 Boxers of All Time, 2017.

[246] Sarah Bartlett. European Cities on a Budget, April 2018.

[247] POLYGRAPH, Alberto Cairo, and Simon Rogers. World Cup 2018: How Every Country is Searching, 2018.

[248] Chris Donaldson, Lee Masters, and Tom Williams. Uncomfortable Questions Concerning Space And Time, 2018.

[249] The Economist. Every World Cup goal ever scored. *The Economist*, June 2018.

[250] Meg Gleason. Cheese is Grate, 2018.

[251] Ashlyn Still, Hang Huang, and Christine Chan. Unpacking Amazon.com and its Prime economy, 2018.

[252] Periscopic. United States gun death data visualization by Periscopic, 2013.

[253] Christine Quan. Airbnb Experiences: An Emoji Story, 2018.

[254] Billy Ker, Chee Wei Xian, and Denise Chong. A who's who guide to the Marvel Cinematic Universe, 2018.

[255] Arthur Soares, Beatrys Rodrigues, Cintia Ferreira, Laura Del Vecchio, Lidia Zuin, Lucas Munhoz, Lucuana dos Anjos, Michell Zappa, Quentin Ladetto, Rafael Pelosini, Rafael Ribeiro, Thiara Cavadas, and Thomaz Rezende. SFSF | Blade Runner 2049, 2017.

[256] The Economist. The Big Mac index. *The Economist*, January 2019.

[257] James Offer. Commonwealth War Dead: First World War Visualised, 2014.

[258] John Nelson. UFO Sightings, June 2015.

[259] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

APPENDICES

APPENDIX A
DEMOGRAPHIC SUMMARY

A.1 Participant Demographic Summary

Below is a demographic summary of all recruited participants in the dissertation. Note that some of the data have been dropped and therefore the numbers below might more than what included in the study reports. In summary. there are in total 343 participants been recruited in this dissertation and its related research projects, 143 of them are females, their average age is 31.14. Below is a summary table of all participants in the four phrases of this dissertation (See Table A.1).

Table A.1.: Participant demographic summary of all participants recruited in the four phrases of this dissertation.

		Participant Demographic Information			
Study Phrase		**Total**	**Female**	**Male**	**Average Age**
Stage 1	Initial exploration	13	8	5	27
	Follow-up	12	7	5	27.92
Stage 2	Tryout Session	7	3	4	NA
	Domain expert	3	1	2	NA
Stage 3	Field Test first round	48	9	39	32.71
	Field Test second round	54	19	35	30.8
	Field Test third round	188	78	109	33.26
Stage 4	Follow-up Field Test	48	18	30	35.15
Grand Total		**373**	**143**	**229**	**31.14**

A.1.1 Stage 1 Demographic Summary

Initial Study There are totally 13 participants in the initial study of stage 1. Below are the participant demographic summaries on gender, age, degree, and major, respectively.

Table A.2.: Participants' gender in the initial study of stage 1.

Gender	
female	8
Male	5
Grand Total	13

Table A.3.: Participants' age in the initial study of stage 1.

Age	
Average	27.00
Min	18
Max	47
Std	9.36

Table A.4.: Participants' age in the initial study of stage 1, grouped by age range.

Age range	
18	1
19-25	8
26-30	1
31-35	1
36-40	0
41-45	0
45-50	2
50 up	0
Grand Total	13

Table A.5.: Participants' highest degree in the initial study of stage 1, grouped by degree type.

Degree	
Associate degree	1
Bachelors degree	6
High school graduate, diploma or the equivalent	2
Masters degree	3
Some college credit, no degree	1

Table A.6.: Participants' major in the initial study of stage 1, grouped by category.

Major	
Agriculture and Related Sciences	2
Communication and Journalism	1
Education	3
Engineering	3
Health Professions and Related Clinical Sciences	1
Other	1
Public Administration and Social Services	2

Follow-up Study There are totally 12 participants in the follow-up study of stage 1. Below are the participant demographic summaries on gender, age, degree, and major, respectively.

Table A.7.: Participants' gender in the follow-up study of stage 1.

Gender	
Female	7
Male	5
Grand Total	12

Table A.8.: Participants' age in the follow-up study of stage 1.

Age	
Average	27.92
Min	21
Max	52
Std	8.95

Table A.9.: Participants' age in the follow-up study of stage 1, grouped by age range.

Age range	
18-25	7
26-30	3
31-35	0
36-40	1
41-45	0
45-50	0
50 up	1
Grand Total	12

Table A.10.: Participants' highest degree in the follow-up study of stage 1, grouped by degree type.

Degree	
Associate degree	1
Bachelors degree	4
High school graduate, diploma or the equivalent	3
Masters degree	2
Some college credit, no degree	2

Table A.11.: Participants' major in the follow-up study of stage 1, grouped by category.

Major	
Agriculture and Related Sciences	1
Business	3
Engineering	5
Liberal Arts, General Studies, and Humanities	2
Other	1

A.1.2 Stage 2 Demographic Summary

Tryout Session There are totally 7 participants in the tryout session of stage 2. Below are the participant demographic summaries on gender, degree, and major, respectively.

Table A.12.: Participants' gender in the tryout session of stage 2.

Gender	
Female	3
Male	4
Grand Total	7

Table A.13.: Participants' highest degree in the tryout session of stage 2, grouped by degree type.

Degree	
High school graduate, diploma or the equivalent	2
Bachelors degree	1
Masters degree	4

Table A.14.: Participants' major in the tryout session of stage 2, grouped by category.

Major	
Computer Science	2
Technology	3
Engineering	1
Politics	1

Expert Review There are totally 3 participants in the expert review of stage 2. Below are the demographic information of three participants. There are 2 male and 1 female, all 3 experts have PhD degree.

Expert no.1

- Occupation: Associate Professor
- Research Domain: Information visualization, visual analytics, human-computer interaction
- Years of Experience: 18 years

Expert no.2

- Occupation: Associate Professor
- Research Domain: HCI, Immersive Systems, Information Visualization
- Years of Experience: 8 years

Expert no.3

- Occupation: Instructor
- Research Domain: Information Visualization, CS Education
- Years of Experience: more than 4 years

A.1.3 Stage 3 Demographic Summary

Field Test Round 1 There are totally 48 participants in the field test round 1 of stage 3. Below are the participant demographic summaries on gender, age, and degree, respectively.

Table A.15.: Participants' gender in the field test round 1 of stage 3.

Gender	
Female	9
Male	39
Grand Total	48

Table A.16.: Participants' age in the field test round 1 of stage 3.

Age	
Average	32.71
Min	24
Max	70
Std	9.32

Table A.17.: Participants' age in the field test round 1 of stage 3, grouped by age range.

Age range	
18-25	10
26-30	14
31-35	12
36-40	6
41-45	1
45-50	2
50 up	3
Grand Total	48

Table A.18.: Participants' highest degree in the field test round 1 of stage 3, grouped by degree type.

Degree	
High school graduate, diploma or the equivalent	4
Professional degree	7
Some college credit, no degree	7
Some high school, no diploma	1
Bachelor's degree	30
Master's degree	4
Doctorate degree	1

Field Test Round 2 There are totally 54 participants in the field test round 2of stage 3. Below are the participant demographic summaries on gender, age, and degree, respectively.

Table A.19.: Participants' gender in the field test round 2 of stage 3.

Gender	
Female	19
Male	35
Grand Total	54

Table A.20.: Participants' age in the field test round 2 of stage 3.

Age	
Average	32.71
Min	24
Max	70
Std	9.32

Table A.21.: Participants' age in the field test round 2 of stage 3, grouped by age range.

Age range	
18-25	12
26-30	22
31-35	11
36-40	4
41-45	1
45-50	2
50 up	2
Grand Total	54

Table A.22.: Participants' highest degree in the field test round 2 of stage 3, grouped by degree type.

Degree	
High school graduate, diploma or the equivalent	7
Professional degree	1
Some college credit no degree	10
Bachelors degree	28
Masters degree	6

Field Test Round 3 There are totally 188 participants in the field test round 3 of stage 3. Below are the participant demographic summaries on gender, age, and degree, respectively.

Table A.23.: Participants' gender in the field test round 3 of stage 3.

Gender	
Female	78
Male	109
Other	1
Grand Total	188

Table A.24.: Participants' age in the field test round 3 of stage 3.

Age	
Average	33.26
Min	20
Max	72
Std	10.63

Table A.25.: Participants' age in the field test round 3 of stage 3, grouped by age range.

Age range	
18-25	39
26-30	64
31-35	34
36-40	21
41-45	5
45-50	10
50 up	15
Grand Total	188

Table A.26.: Participants' highest degree in the field test round 3 of stage 3, grouped by degree type.

Degree	
High school graduate, diploma or the equivalent	16
Professional degree	1
Some college credit, no degree	23
Associate degree	13
Bachelor's degree	114
Master's degree	16
Doctorate degree	2
Trade/technical/vocational training	3

A.1.4 Stage 4 Demographic Summary

Follow-up Field Test There are totally 48 participants in the follow-up field test of stage 4. Below are the participant demographic summaries on gender, age, and degree, respectively.

Table A.27.: Participants' gender in the follow-up field test of stage 4.

Gender	
Female	18
Male	30
Grand Total	48

Table A.28.: Participants' age in the follow-up field test of stage 4.

Age	
Average	35.15
Min	23
Max	62
Std	9.92

Table A.29.: Participants' age in the follow-up field test of stage 4, grouped by age range.

Age range	
18-25	9
26-30	10
31-35	9
36-40	8
41-45	5
45-50	2
50 up	5
Grand Total	48

Table A.30.: Participants' highest degree in the follow-up field test of stage 4, grouped by degree type.

Degree	
High school graduate, diploma or the equivalent	3
Some college credit, no degree	5
Associate degree	6
Bachelor's degree	27
Master's degree	5
Trade/technical/vocational training	2

APPENDIX B
EXPERIMENT MATERIALS

B.1 Experiment Materials in Pilot Test

B.1.1 Tested Visualizations

Table B.1.: The name, size, and interaction level of the three visualizations used in the pilot study.

Name	Size	Interaction
Record Breaking Roller-coaster [228]	partial screen	Advanced
Analysis of Queen [229]	partial screen	basic
Women in Tech [230]	partial screen	static

B.2 Experiment Materials in Stage 1

B.2.1 Tobii Studio Recording Video Export Setup

Tobii Studio 3.4.8 [142] is used for the study. To overlay mouse and eye tracking trace on the screen recording with user sound, first, go to the drop-down menu, enter Tools >Settings. When the Global Settings window is opened, go to "Screen and Video Capture" Tab, make "Frame rate" value under "Screen Capture" as 30 or more, and check "Record User Sound" option under "Audio". Finally, click "ok" to complete (See Figure B.1).

Then, conduct eye-tracking and record the session with Tobii Studio as usual.

When the recording is complete, go to the "Replay" tab in the main window, select one of the recording and click "Export Movie" button on the top menu, t to

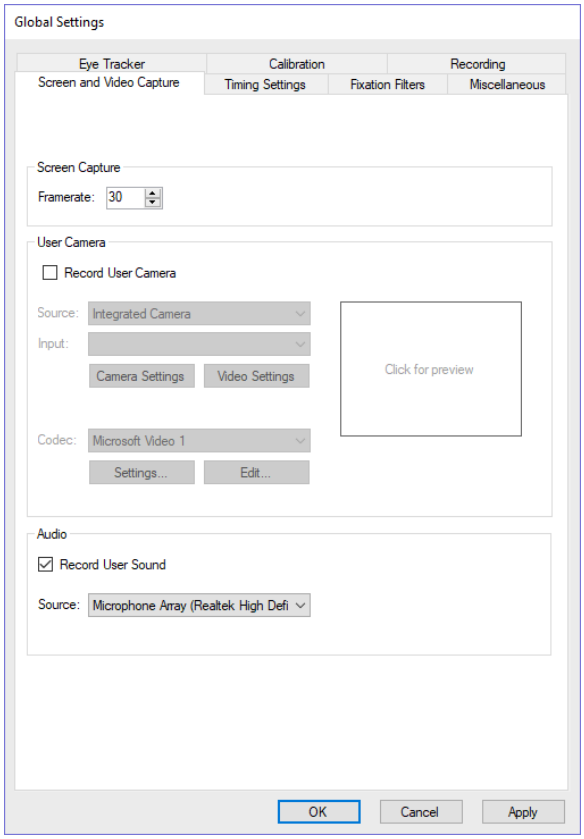


Figure B.1. Screenshot of "Screen and Video Capture" tab in "Global Settings" window.

open the window "Batch Export Segments to AVI Clips". After setup the "Export folder", check any of the option under "Export Recordings. In "AVI encoder", edit the Video codec as "Microsoft Video 1", and make "Frame rate" value as 30 or more (See Figure B.2). Finally, click "ok". Wait for a while and save the exported .avi files separately.

B.2.2 Tested Visualizations in Rating Session

Before the think-aloud sessions, there is a rating session where participants saw the screenshots of 10 visualizations and provided their level of interest (See Table B.2). The instructions of grading activities are list below:

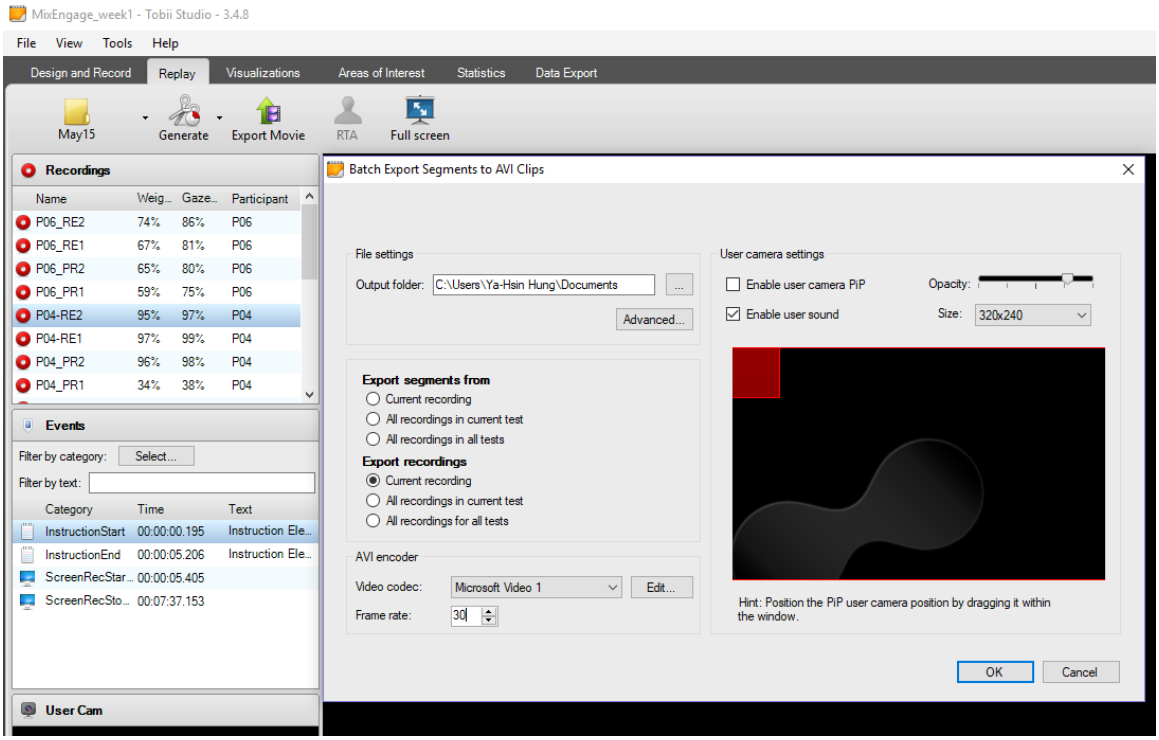


Figure B.2. Screenshot of "Batch Export Segments to AVI Clips" window.

Grade 10 Diagrams

Please go through the following 10 diagrams screenshots and grade them based on your level of interest. Your level of Interest (1 to 10):

B.2.3 Tested Visualizations in Practice Session

There are two practice sessions for participants to get familiar with the actual think-aloud session. Two visualizations were used for the practice purposes (see Table B.3).

Table B.2.: The name, size, and interaction level of the ten visualizations used in grading sessions of the experiment

#	**Name**	**Size**	**Interaction**
A	Hierarchical edge bundling [231]	partial screen	basic
B	70 years of first names in France [232]	partial screen	basic
C	Gapminder Tools [233]	full screen	advanced
D	Wealth Inequality in the US [234]	full screen	basic
E	35 Years Of American Death [235]	more than one screen	basic
F	U.S. Presidents [147]	full screen	advanced
G	2015 NFL Predictions [236]	partial screen	basic
H	Oscar Predictions [148]	more than one screen	basic
I	Medals Won by Olympic Athletes [237]	full screen	advanced
J	New York Taxis [238]	full screen	basic

Table B.3.: The name, size, and interaction level of the two visualizations used in practice sessions.

Name	**Size**	**Interaction**
When to eat chocolate: A guide for Researchers [149]	partial screen	static
Causes of Death [150]	partial screen	interactive

B.2.4 Tested Visualizations in Think-aloud Session

For each participant, there were two think-aloud sessions. All 12 participants were assigned the same two visualizations in the two sessions with the order was randomized.

Table B.4.: The name, size, and interaction level of the two visualizations used in the think-aloud sessions.

Name	**Size**	**Interaction**
U.S. Presidents [147]	full screen	advanced
Oscar Predictions [148]	more than 1 screen	basic

B.2.5 Tested Visualizations in Follow-up Study

In the follow-up study, every participants received different visualizations in their think-aloud sessions. The assignment of the tested visualizations can be found in table B.5.

Table B.5.: The assignment of tested visualizations in the 1st and the 2nd think-aloud sessions for the follow-up study.

#	**1st think-aloud session**	**2nd think-aloud session**
P14	US Household Income Distribution	Dam Nation: State of US Dams
P15	The History of the US	Offenses in Ivy League Schools
P16	The New York Times Best Seller List	The Video Game Console War
P17	Major League Soccer Attendance	Surnames and Race in the U.S
P18	The Price of Curry	How Much Cash Does a Freshman Generate?
P19	Billionaires by Forbes	Travel Visa inequality
P20	OECD Regional Well-Being	The Timing of Baby Making
P21	What city is the microbrew capital of the US?	The Shape of Slavery
P22	Most Unlikely Comebacks	Global Gender Gap Report Browser
P23	Histography	How Ebola Spreads
P24	Commonwealth War Dead: First World War Visualised	Resourcetrade.earth
P25	Job Market Tracker	An Ocean of Noise

B.2.6 Post-Trial Survey Questionnaires

Demographic survey

1. What is your birth year?
2. What is your gender?
   (a) Male
   (b) Female
3. What is the highest degree or level of school you have completed? If currently enrolled, select the highest degree received.
   (a) No schooling completed
   (b) Nursery school to 8th grade
   (c) Some high school, no diploma
   (d) High school graduate, diploma or the equivalent (for example: GED)
   (e) Some college credit, no degree
   (f) Trade/technical/vocational training
   (g) Associate degree
   (h) Bachelors degree
   (i) Masters degree
   (j) Professional degree
   (k) Doctorate degree
4. What is your field of work or study?
   (a) Agriculture and Related Sciences
   (b) Arts, Visual, and Performing
   (c) Business
   (d) Communication and Journalism
   (e) Computer and Information Sciences
   (f) Education
   (g) Engineering
   (h) Health Professions and Related Clinical Sciences
   (i) Law
   (j) Social Sciences

(k) Economics

(l) Liberal Arts, General Studies, and Humanities

(m) Public Administration and Social Services

(n) Sciences and Math

(o) Others

22 User Engagement Assessment Items The items using slider selection tool in Qualtrics, the sequence will be randomized.

Please answering following questions based on your experience during this trial. (Strong Agree, Agree, Slightly Agree, Neutral, Slightly Disagree, Disagree, Strong Disagree)

1. While using this interactive chart, I found its look and feel to be pleasing.
2. The layout of this interactive chart is clear and balanced.
3. While using this interactive chart, I felt absorbed to the extent that I was not aware of my surroundings.
4. While using this interactive chart, time seemed to pass quickly.
5. While using this interactive chart, I enjoyed and accepted any challenges it presented.
6. While using this interactive chart, I had to think carefully, deeply, or reflectively.
7. While using this interactive chart, its functions and features worked as I expected.
8. While using this interactive chart, I felt in control.
9. While using this interactive chart, I learned something that I had not known before (e.g., a new fact, concept, or piece of information).
10. While using this interactive chart, I learned and figured out how to use it along the way.
11. While using this interactive chart, I felt as though I was moving in or through it to learn about its content or message.

12. While using this interactive chart, I was exploring its features and content in a gradual fashion.
13. While using this interactive chart, I found myself imagining things not directly related to what I was seeing in the chart.
14. While using this interactive chart, I found myself generating new and original thoughts or ideas.
15. While using this interactive chart, I found myself concentrating on specific aspects or features of the chart.
16. While using this interactive chart, I had to pay attention to multiple things at the same time.
17. The content or message of this interactive chart was interesting to me.
18. The features or interactions provided in this interactive chart were interesting to me.
19. The look and feel of this interactive chart was novel and fresh.
20. The features or interactions provided in this interactive chart were novel and fresh.
21. While using this interactive chart, I experienced enjoyment from the chart in and of itself, and not because it was a means to an end.
22. I would want to use this interactive chart if I saw it somewhere else and was not required or encouraged to use it.
23. (Open-ended Question) conducting this experiment, had you seen this interactive chart? What aspect(s) of this interactive chart did you find most engaging?

B.3 Experiment Materials in Stage 2

B.3.1 Initial Item Bank

Below is the initial item bank which compromised 137 item candidates. Items are grouped by the indicator and categorized based on three categories followed by the corresponding indicators.

Behavior Level

Fluidity Continuity; Un-Disruption; the flow of use can't be stopped; the flow of use should be continues

- There is no interruption during my use
- I feel my use of this chart is continuous
- I don't want to be interrupted when I use this chart
- I don't want to be stopped when I use this chart
- I want to completely explore this chart

Enthusiasm Being enthusiastic; Intriguing; really into something, put a lot of effort in it

- I want to see more charts and diagram like this
- I am motivated to know more about this topic
- I am curious about the motivation
- Feel curious about back story of this visualization
- (Reverse) Doubt about the data of this visualization
- (Reverse) Have concerns about the information/message in this visualization
- Have opinions about the information/message in this visualization
- I have personal opinion
- I put effort into exploring this visualization willingly

Curiosity Feeling of exploration. Sense of wonder.

- Enjoy when interacting with the visualization
- I feel I have a good time to see everything in this chart
- I enjoy checking the details in this visualization
- I enjoy wandering around in this visualization
- Desire to know something
- I am willing to explore this chart the more I use it
- I am willing to take longer time to explore this visualization thoroughly
- Be curious about something
- I keep asking myself questions when exploring this visualization
- Curiosity is my driving forces when I explore this visualization
- Try to speculate based on the investigation
- I tried to speculate based on my investigation
- I have speculated about the content of the visualization
- I speculated about the visuals of the visualization
- I speculated about the features/interactions of the visualization

Discovery A process or a phenomena where users find out new things along the way.

- Learn something that not known before (e.g., a new fact, concept, or piece of information).
- I feel I have learned something that not known before (e.g., a new fact, concept, or piece of information)
- I gradually figured out how to use the visualization the more I use it
- Notice something that can raise awareness
- Feel gain something from the chart
- I acquired (e.g., a new fact, concept, or piece of information) something useful when using this visualization
- Be self-motivated to learn more about the visualization
- I Feel exciting to know whats next in the visualization

Search Part of "Discovery". Feel moving in or through the visualization (e.g., topic, message).

- Explore the visualization in a gradual fashion.
- Learn more and more on the visualization along the way
- There are multiple stages during the exploration
- Feel go through several layers/stages of information/messages within the visualization
- Bring the information learned from the previous section to the next one
- Information learned from the previous section motivates the following exploration
- information learned from the previous section helps on understanding the next one
- (Reverse) The exploration is not successful
- (Reverse) Feel lost when explore this visualization
- (Reverse) Feel lost when I try to learn more on the visualization
- (Reverse) I felt uncertain when I explore this visualization

Judgment Level

Clarity Expressiveness; how well this visualization shows/expresses its concept/message. How well this visualization presents information in a clear and complete fashion.

- Can feel the richness of the information
- The chart clearly expresses a certain message
- The chart clearly present a certain concept
- The visualization effectively deliver its main concept and idea
- The chart deliver a clear (positive/negative) emotion

Persuasion Part of "Clarity"; agree with the points conveyed by the visualization.

- The impact or influence on the viewers; how powerful the message or impact this viz can generated

- I agree with the concept or the message contained in this chart
- I think the visualization persuade me on this topic
- The message / information in this visualization is persuasive
- I think I believe the content information in this visualization
- I will agree with the message in this visualization
- I feel the messages in this chart inline with my values
- I believe this visualization can be influential to its readers
- I am touched because of this chart
- I think this visualization influences my emotion
- Feel the message is powerful in this visualization

Triggering Part of "Captivation"; Some positive elements (e.g., imagination, pleasure) are provided from the use of visualization, and therefore triggered further exploration. A positive loop.

- I feel I got some rewards from using this chart
- The gained knowledge itself is the rewards of using this visualization
- The enjoyment is the rewards of using this chart
- The novelty is the rewards of using this chart
- The feeling of discover something is the rewards of using this visualization
- I feel amazed by the visualization

Storytelling How well the information been delivered to the users.

- The information provided in this visualization is easy to comprehend
- Feel the way information been organized is intriguing
- Think the information been illustrated in this visualization is interesting
- Think the information been illustrated in this visualization is memorable
- The information provided in this visualization impressed me

Creativity Creative and original thoughts have been generated.

- Imagine things not directly related to what can be seen in the visualization.

- I think this visualization triggers my imaginations outside of its topic
- I feel this visualization is inspiring
- I think this visualization triggers my creative thoughts
- Generate new and original thoughts
- I feel I have some creative thoughts after using this visualization
- I think this visualization clicks something in my head
- Thinking about related information
- Triggering more advanced (insightful) thoughts
- Generate associated ideas
- Promote creative thoughts

Feeling Level

Entertaining Interesting; a reaction shows that this visualization is fun or likable.

- I feel using this visualization is entertaining
- Feel the topic of the visualization is entertaining
- Feel the way the visualization presents information is entertaining
- I feel entertained when using this visualization
- Feel the story in the visualization is entertaining
- Feel relaxing when using the visualization
- Feel the content (e.g., topic, message) of the visualization is interesting.
- I feel the topic of this visualization is interesting to me
- I feel the data provided in this visualization is interesting to me
- The interpretation of this visualization is interesting to me
- The implication of this visualization is interesting to me
- I feel I acquire something interesting in this visualization
- Feel the features (e.g., interactions, animations) provided in the visualization is interesting.
- I enjoyed the moment of using this visualization

- I have a positive impression on this visualization
- I would like to recommend this visualization to my friends

Untroubling Being Controlled; a situation where use feels in control when using the visualization

- The visualization is worked as expected (e.g., functions and features).
- I feel in control when using the visualization
- Work as expected
- The chart provides expected feedback
- Be able to anticipate the next move
- The use(control) is smooth
- My control on this visualization is effortless
- I feel no difficulty when using this visualization
- Feel comfortable when controlling the visualization
- (Reverse) Feel confused at some points
- (Reverse) I feel confused when figuring out how to use this visualization
- (Reverse) Be annoyed by the control of this visualization
- (Reverse) Feel helpless when using this visualization

Novelty Part of "Creativity"; Feel target visualization (one aspect) is novel.

- I feel the look of the visualization is novel.
- I feel the features (e.g., interactions) in the visualization are novel
- I feel the data and information in the visualization is novel
- The visual style of this visualization looks novel to me
- I feel the way this visualization shows its data is new to me
- I feel the media of showing the data is novel
- The novelty of this visualization attracts me

Captivation Been captivated or absorbed by the visualization.

- I feel being absorbed to the extent

- I feel the time seems to pass quickly
- I forget about time when using
- I was not aware of the surroundings during my use of this visualization

Pleasing Aesthetics-wise aspects, balance visually, have plenty of stuff, give a feeling of satisfaction.

- I feel the visualization looks to be pleasing
- Feel the visualization looks rich
- Think the layout of the visualization is clear
- I feel the design style of the visualization is pleasing
- I am impressed by the design style of the visualization
- Feel the color use of the chart is pleasing
- Feel the way visualization presents the information is clever
- Feel the visual design of the visualization is nice
- Can feel the bounty from the visuals
- The visualization demonstrate a beautiful view
- The chart clearly shows a vivid world

B.3.2 Item Bank After Tryout Sessions

Below is the item bank which compromised 80 items after being review by intended respondents of AEVis. Items are grouped by the indicator and categorized based on three categories: Behavior (see Table B.6), judgment (see Table B.7), and feeling (see Table B.8).

Table B.6.: Items under behavior category, grouped by the corresponding indicator: Fluidity, Enthusiasm, Curiosity, and Discovery.

Behavior (user's specifc behavioral pattern)	
Fluidity	There am focused during my use of this visualization
	I want to be focused when I use this visualization
	I want to keep going when using this visualization.
	My use of this visualization is continuous and smooth.
	My use of this visualization is fluid.
	The flow of use for this visualization is continuous.
Enthusiasm	I am eager to proceed when using this visualization.
	I feel motivated while using this visualization
	I put effort into exploring this visualization
	I am enthusiastic when using this visualization
	I want to learn more when I use this visualization
	I am proactive when using this visualization
Curiosity	I keep inspecting this visualization
	I enjoy exploring this visualization
	I was willing to take time to explore this visualization.
	I was willing to explore this visualization thoroughly
	I maintain curiosity when using this visualization
	I keep speculating and investigating when using this visualization
Discovery	I feel I have learned a new concept or new knowledge from the visualization
	I have acquired a new concept or new knowledge from the visualization
	I want to know more about this topic after using this visualization
	The knowledge I gained is a sufficient reward of using this visualization
	The feeling of discovery is the reward of using this visualization
	I learn more and more about the visualization as explore

Table B.7.: Items under judgment category, grouped by the corresponding indicator: Clarity, Storytelling, and Creativity.

Judgment (user's judgment of the visualization)	
Clarity	I feel the visualization clearly expresses a certain information
	I feel the visualization clearly present a certain concept
	I think the visualization effectively delivers its main concept or idea
	I think the visualization clearly delivers a point of view
Storytelling	I think this visualization is easy to comprehend
	I feel this visualization is communicating a good story or telling a good point
	I think the content of this visualization is interesting
	I think this visualization is memorable
	I think this visualization is impressing
	I think this visualization is persuasive
	I think this visualization is telling a compelling story
	I believe this visualization can be influential to its readers
	I think this visualization influences my emotion
	I understand what the visualization is telling
Creativity	I think this visualization sparks my imagination
	I think this visualization is inspiring
	I think this visualization sparks my creative thinking
	I came up with some creative thoughts after using this visualization
	I feel the look of the visualization is novel.
	The visual of this visualization looks novel to me
	The way this visualization shows its data is new to me
	The novelty of this visualization attracts me
	I think a visualization initiates my creative thoughts/idea

Table B.8.: Items under feeling category, grouped by the corresponding indicator: Entertainment, Untroubling, Captivation, and Pleasing.

Feeling (user's reaction of the visualization)	
Entertainment	Using this visualization is entertaining
	I feel entertained when using this visualization
	I feel the content (e.g., topic, message) of the visualization is interesting.
	Using this visualization is insteresting
	I find entertainting when using the visualization
	I am interested when using this visualization.
	Using this visualization is enjoyable.
	I find this visualization is enjoyable when using it.
	I find myself enjoyed using this visualization.
Untroubling	I feel in control when using the visualization
	The visualization provides expected feedback
	Be able to anticipate the next move
	My control on this visualization is effortless
	I feel no difficulty when using this visualization
	(Reverse coded) I feel confused when figuring out how to use this visualization
	(Reverse coded) Be annoyed by the control of this visualization
	(Reverse coded) Feel helpless when using this visualization
	(Reverse coded) I feel I was lost when exploring this visualization
	(reverse coded) I feel frustrated when using this visualization
	I do not feel frustrated when using this visualization
Captivation	I feel absorbed by this visualization while using it
	I feel the time seems to pass quickly when using this visuslization
	I forgot about time when using this visualization
	I was not aware of the surroundings when using this visualization
	I was concentrated on the visualization when using it
	I was not aware of time when using this visualization
	I feel captivated while using this visualization
Pleasing	The look and feel of the visualization is pleasing to me
	The design style of the visualization is pleasing
	I am impressed by the design of this visualization
	I am pleased by the design of this visualization
	Feel the design of the visualization is good
	The look and feel of this visualization is attractive to me.

B.3.3 Expert Review Survey

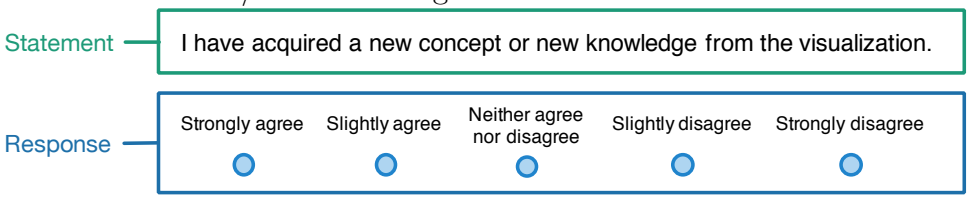
Below is an example of the online survey form used in the expert review. Each expert would get different sets of items to review.

Content Validity: Affective Engagement

An important phase in the development of any instrument is that of content validation. By offering your expertise, you are contributing to the development of an instrument that is content valid. Your assistance in this phase of instrument development is sincerely appreciated. Thanks in advance for your time and help. (IRB Information sheet)

Affective Engagement (AE) in this study refers to a construct dealing with a user's emotional and mental involvement, attraction, fascination, and captivation when he or she interacts with an information visualization. There are 11 elements we identified as indicators of affective engagement (where did these indicators come from?):

The following figure is an example question in the final survey. The respondent have to answer his/her level of agreement of the statement.



In the following pages, we have prepared 11 statement that can be used to assess the level of its corresponding indicator, each of them is being considered for inclusion in a new survey questionnaire for measuring affective engagement in information visualization. You will be providing three ratings for each item. The rating tasks are listed below.

Figure B.3. General introduction of expert review survey.

Indicators	Description
Fluidity	The flow of use is continuous, smooth, not disrupted
Enthusiasm	The user is interested, eager, proactively involved, and motivated
Curiosity	The user is inquisitive or investigative
Discovery	The user discovers something new or noticeable
Clarity	The visualization express a clear concept or message; the data is clear or easily interpreted
Storytelling	The visualization tells a compelling story or has a persuasive narrative style
Creativity	The visualization helps users generate or express creative and innovative thoughts or ideas
Entertaining	The user feels the visualization is fun, interesting, or charming
Untroubled	The user feels content and doesn't feel upset or frustrated
Captivation	The user is concentrating, absorded
Pleasing	The user is impressed by the visualization (e.g., visually, concept-wise)

Figure B.4. The table contains 11 AE indicators and the descriptions for each of them.

Rating Tasks You will rate each item stem (statement of each question) on the following aspects:

- **A. Clarity** Please indicate your perception of the clarity of the item (conciseness, readability, understandability)
- **B. Relevance** Please indicate how relevant you feel each item is to the construct of affective engagement.
- **C. Comments** Feel free to write comments regarding the item stems. (These comments could suggest changes in wording or suggest that the item be eliminated).

Figure B.5. Rating Task instruction of expert review survey

Demographic information

1. Your current professional title:
   - Faculty
   - Student
   - Research scientist
   - Other
2. Your research area(s):
3. Years of experience in visualization:

Figure B.6. Expert Demographic survey

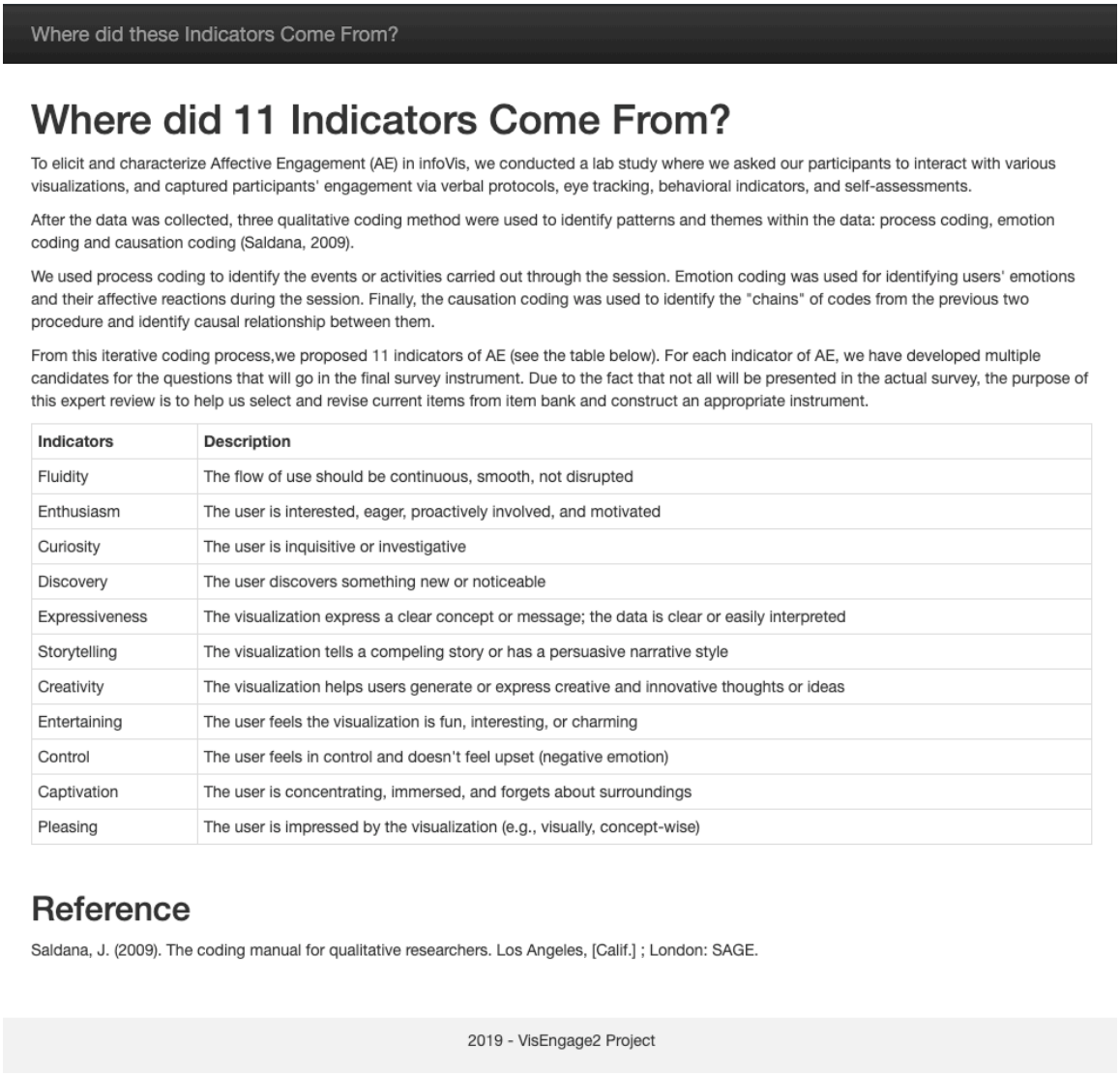


Figure B.7. Additional webpage for domain experts to understand the overview of the study and where the 11 indicators came from.

B.4 Experiment Materials in Stage 3

B.4.1 Tested Visualizations

Candidates of tested visualizations in field test are listed below.

Table B.9.: Name, size, and interaction type of tested visualizations in field test

Name	Size	Interaction
The Evolution of the Office Desk [239]	more than 1 screen	static
Sugar: The Bitter Truth [240]	more than 1 screen	static
Avengers Assemble [241]	full screen	static
Say it in Engrish [242]	more than 1 screen	static
The Carbon Footprint of the Internet [243]	full screen	static
The 100 greatest films of the Century [244]	more than 1 screen	static
FIFA World Cup Tournaments [170]	partial screen	basic
Hard Knuckles - Top 50 Boxers Of All Time [245]	more than 1 screen	basic
European Cities on a Budget [246]	full screen	basic
World Cup 2018: How the World is Searching [247]	full screen	advanced
Uncomfortable Questions Concerning Space And Time [248]	full screen	static
Every World-Cup Goal Ever Scored [249]	partial screen	advanced
Cheese is Grate [250]	full screen	static
Prime Economy [251]	more than 1 screen	basic
U.S. Gun Deaths [252]	partial screen	basic
Airtweets: An Emoji Story [253]	more than 1 screen	basic
A Who's Who Guide to the Marvel Cinematic Universe [254]	full screen	advanced
2001: A Space Odyssey [171]	full screen	advanced
Blade Runner 2049 [255]	full screen	advanced
The Big Mac index [256]	partial screen	advanced
Commonwealth War Dead: First World War Visualised [257]	more than 1 screen	basic
UFO Sightings [258]	more than 1 screen	static
Creative Routines [169]	more than 1 screen	static
Most FIFA World Cup goals	partial screen	basic

Content-related Questions

Content-related questions for visualization no.1.

1. What this infographic is about? (multiple-selection)
   - Career of famous writers
   - Income of famous musicians
   - lifespan of famous scientists
   - Daily routine of creative people ✓
2. How many hours does Benjamin Franklin spend on sleeping? (multiple-selection)
   - 6
   - 7 ✓
   - 8
   - 10
3. How many creative people are listed in this infographic? (short answer)
   - Answer: 16

Content-related questions for visualization no.2 (original).

1. What this interactive chart is about? (multiple-selection)
   - Europa league
   - FIFA World Cup ✓
   - Champion league
   - World Championship
2. Which team faced France in the 2018 FIFA World Cup quarter final? (multiple-selection)
   - England
   - Brazil
   - Croatia
   - Uruguay ✓

3. How many goals did Croatia score in 2018 FIFA World Cup final? (short answer)
    - Answer: 2

Content-related questions for visualization no.2 (modified).

1. What this interactive chart is about? (multiple-selection)
    - Europa league
    - FIFA World Cup ✓
    - Champion league
    - World Championship
2. Which country has the most goals in 2010's tournament? (multiple-selection)
    - France ✓
    - Brazil
    - Germany
    - Spain
3. How many goals did Italy score in 2002's tournament? (short answer)
    - Answer: 11

Content-related questions for visualization no.3.

1. What this interactive chart is about? (multiple-selection)
    - Technologies used in "Ghost in the Shell"
    - Technologies used in "Blade Runner 2049"
    - Technologies used in "2001: A space Odyssey" ✓
    - Technologies used in "Aliens"
2. Technology that numbered as 35 is? (multiple-selection)
    - Paper thin screen
    - Garbage drone
    - Jetpack
    - Sleeping pod ✓

3. The number that assigned to technology "Speech Translator" is? (short answer)
   - Answer: 27

B.4.2 Demographic survey

1. What is your birth year?
2. What is your gender?
   (a) Male
   (b) Female
3. What is the highest degree or level of school you have completed? If currently enrolled, select the highest degree received.
   (a) No schooling completed
   (b) Nursery school to 8th grade
   (c) Some high school, no diploma
   (d) High school graduate, diploma or the equivalent (for example: GED)
   (e) Some college credit, no degree
   (f) Trade/technical/vocational training
   (g) Associate degree
   (h) Bachelors degree
   (i) Masters degree
   (j) Professional degree
   (k) Doctorate degree
4. What is your field of work or study?
   (a) Agriculture and Related Sciences
   (b) Arts, Visual, and Performing
   (c) Business
   (d) Communication and Journalism
   (e) Computer and Information Sciences

(f) Education

(g) Engineering

(h) Health Professions and Related Clinical Sciences

(i) Law

(j) Social Sciences

(k) Economics

(l) Liberal Arts, General Studies, and Humanities

(m) Public Administration and Social Services

(n) Sciences and Math

(o) Others

APPENDIX C
ANALYSIS CODES

C.1 R Codes for Statistical Analysis

The R syntx for item analysis, EFA, and CFA that used in stage 3, R version 3.5.3 [259].

C.1.1 Item Analysis

```
# Require packages
    install.packages("polycor")
    install.packages("psych")
    require(polycor)
    require(psych)
# basic descriptive statistics.
    summary(all_data)
# unformatted frequency tables
    lapply(all_data, table)
# variability index (standard deviation)
    lapply(all_data, sd)
############ polyserial correlation #############
#---------------behavior-----------------------------
    polyserial(TotalSum, Q1, std.err = FALSE, ML = TRUE)
    polyserial(BehvSum, Q1, std.err = FALSE, ML = TRUE)
    polyserial(TotalSum, Q2, std.err = FALSE, ML = TRUE)
    polyserial(BehvSum, Q2, std.err = FALSE, ML = TRUE)
```

```
    polyserial(TotalSum, Q3, std.err = FALSE, ML = TRUE)
    polyserial(BehvSum, Q3, std.err = FALSE, ML = TRUE)
    polyserial(TotalSum, Q4, std.err = FALSE, ML = TRUE)
    polyserial(BehvSum, Q4, std.err = FALSE, ML = TRUE)
#--------------judgment------------------------------
    polyserial(TotalSum, Q5, std.err = FALSE, ML = TRUE)
    polyserial(JudgSum, Q5, std.err = FALSE, ML = TRUE)
    polyserial(TotalSum, Q6, std.err = FALSE, ML = TRUE)
    polyserial(JudgSum, Q6, std.err = FALSE, ML = TRUE)
    polyserial(TotalSum, Q7, std.err = FALSE, ML = TRUE)
    polyserial(JudgSum, Q7, std.err = FALSE, ML = TRUE)
#---------------feeling------------------------------
    polyserial(TotalSum, Q8, std.err = FALSE, ML = TRUE)
    polyserial(FeelSum, Q8, std.err = FALSE, ML = TRUE)
    polyserial(TotalSum, Q9, std.err = FALSE, ML = TRUE)
    polyserial(FeelSum, Q9, std.err = FALSE, ML = TRUE)
    polyserial(TotalSum, Q10, std.err = FALSE, ML = TRUE)
    polyserial(FeelSum, Q10, std.err = FALSE, ML = TRUE)
    polyserial(TotalSum, Q11, std.err = FALSE, ML = TRUE)
    polyserial(FeelSum, Q11, std.err = FALSE, ML = TRUE)
#---------------Q9 Candidates------------------------
    polyserial(TotalSum, Q9v1, std.err = FALSE, ML = TRUE)
    polyserial(TotalSumv1, Q9v1, std.err = FALSE, ML = TRUE)
    polyserial(TotalSum, Q9v2, std.err = FALSE, ML = TRUE)
    polyserial(TotalSumv2, Q9v2, std.err = FALSE, ML = TRUE)
    polyserial(TotalSum, Q9v3, std.err = FALSE, ML = TRUE)
    polyserial(TotalSumv3, Q9v3, std.err = FALSE, ML = TRUE)
############ pattern of average total test scores ######
    describeBy(all_data$TotalSum,all_data$Q1)
```

```
    describeBy(all_data$TotalSum,all_data$Q2)
    describeBy(all_data$TotalSum,all_data$Q3)
    describeBy(all_data$TotalSum,all_data$Q4)
    describeBy(all_data$TotalSum,all_data$Q5)
    describeBy(all_data$TotalSum,all_data$Q6)
    describeBy(all_data$TotalSum,all_data$Q7)
    describeBy(all_data$TotalSum,all_data$Q8)
    describeBy(all_data$TotalSum,all_data$Q9)
    describeBy(all_data$TotalSum,all_data$Q10)
    describeBy(all_data$TotalSum,all_data$Q11)
```

C.1.2 Exploratory Factor Analysis

```
# Required package
    install.packages("psych")
    install.packages("GPArotation")
    install.packages("semPlot")
    require(psych)
    require(GPArotation)
    library(semPlot)
# basic descriptive statistics
    describe(All_data)
# correlation matrix
    lowerCor(All_data)
# exploratory Factor Analysis
#nfactors=1
    EFAfit1 <- fa(r = All_data, nfactors = 1, rotate = "promax",
    fm = "ml")
#nfactors=2
```

```
    EFAfit2 <- fa(r = All_data, nfactors = 2, rotate = "promax",
    fm = "ml")
#nfactors=3
    EFAfit3 <- fa(r = All_data, nfactors = 3, rotate = "promax",
    fm = "ml")
#nfactors=4
    EFAfit4 <- fa(r = All_data, nfactors = 4, rotate = "promax",
    fm = "ml")
# coefficient omega (and alpha)
#nfactors=1
    omega(All_data, nfactors = 1, fm = "ml", poly = TRUE)
    omega(All_data, nfactors = 1, rotate = "promax", fm = "ml")
    omega1 <- omega(All_data)
    summary(omega1)
#nfactors=2
    omega(All_data, nfactors = 2, fm = "ml", poly = TRUE)
    omega(All_data, nfactors = 2, rotate = "promax", fm = "ml")
    omega2 <- omega(All_data)
    summary(omega2)
#nfactors=3
    omega(All_data, nfactors = 3, fm = "ml", poly = TRUE)
    omega(All_data, nfactors = 3, rotate = "promax", fm = "ml")
    omega3 <- omega(All_data)
    summary(omega3)
#nfactors=4
    omega(All_data, nfactors = 4, fm = "ml", poly = TRUE)
    omega(All_data, nfactors = 4, rotate = "promax", fm = "ml")
    omega4 <- omega(All_data)
    summary(omega4)
```

```
# Draw graph using Semtool
    semPaths(EFAfit2, layout="tree3", whatLabels="std",
    style="lisrel",     edge.color=c("black"),
    color=c("white"), nDigits=3)
```

C.1.3 Confirmatory Factor Analysis

```
# Required package
    install.packages("lavaan")
    library(lavaan)
    library(psych)
    install.packages("semPlot")
    library(semPlot)
# Check that data imported correctly
    dim(All_data)
    View(All_data)
    head(All_data)
################ 3-level model ################
# Specify the model
    TLevel.model <- 'behavior =~ Fluidity + Enthusiasm +
    Curiosity + Discovery
    judgment =~ Clarity + Storytelling + Creativity
    feeling   =~ Entertainment + Untroubling + Captivation +
    Pleasing'
# fit the model
    TLevelfit <- cfa(TLevel.model, data=All_data)
    diagram(TLevelfit)
    # display summary output
    summary(TLevelfit, fit.measures=TRUE,
```

```
    standardized=TRUE, rsquare=TRUE)
# Draw graph using Semtool
    #std statement can be: "no", "est", or "cons"
    semPaths(TLevelfit, layout="tree3", whatLabels="std",
    style="lisrel", edge.color=c("black"), color=c("white"),
    nDigits=3)
############### 2-factor model ###############
# specify the model
    TwoFactor.model <- ' pragmatic  =~ Fluidity + Curiosity +
    Clarity + Untroubling + Pleasing
    hedonic =~ Enthusiasm + Discovery + Storytelling + Creativity +
    Entertainment + Captivation'
# fit the model
    TwoFactorfit <- cfa(TwoFactor.model, data=All_data)
    diagram(TwoFactorfit)
    # display summary output
    summary(TwoFactorfit, fit.measures=TRUE,
    standardized=TRUE,
    rsquare=TRUE)
# Draw graph using Semtool
    #std statement can be: "no", "est", or "cons"
    semPaths(TwoFactorfit, layout="tree3", whatLabels="std",
    style="lisrel", edge.color=c("black"), color=c("white"),
    nDigits=3)
```

C.2 IRTPRO Codes for Statistical Analysis

The .irtpro syntax for IRT analysis in stage 3, IRTPRO 4.2 student version [215].

C.2.1 UniDim GRM

Uni-Dimensional Graded Response Model.

```
Project:
    Name = field test2_1D;
Data:
    File = .\field test2_1D.ssig;
Analysis:
    Name = Test1;
    Mode = Calibration;
Title: Uni-Dimensional Graded Response Model
Comments: Uni-Dimensional Graded Response Model
Estimation:
    Method = BAEM;
    E-Step = 500, 1e-005;
    SE = S-EM;
    M-Step = 50, 1e-006;
    Quadrature = 49, 6;
    SEM = 0.001;
    SS = 1e-005;
Scoring:
    Mean = 0;
    SD = 1;
Miscellaneous:
    Decimal = 2;
    Processor = 1;
    Print M2, CTLD, Loadings, P-Nums, Diagnostic;
    Min Exp = 1;
Groups:
```

```
Group :
    Dimension = 1;
    Items = all_Q1, all_Q2, all_Q9v3, all_Q3, all_Q4, all_Q5,
    all_Q6, all_Q7,
    all_Q8, all_Q10, all_Q11;
    Codes(all_Q1) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q2) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q9v3) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q3) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q4) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q5) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q6) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q7) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q8) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q10) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q11) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Model(all_Q1) = Graded;
    Model(all_Q2) = Graded;
    Model(all_Q9v3) = Graded;
    Model(all_Q3) = Graded;
    Model(all_Q4) = Graded;
    Model(all_Q5) = Graded;
    Model(all_Q6) = Graded;
    Model(all_Q7) = Graded;
    Model(all_Q8) = Graded;
    Model(all_Q10) = Graded;
    Model(all_Q11) = Graded;
    Mean = Free;
    Covariance = Free;
```

```
Constraints:
```

C.2.2 2Dim GRM-Confirm

2-Dimensional Graded Response Model with Full Constraint.

```
Project:
    Name = 2D_GRM_induce;
Data:
    File = .\2D_GRM_induce.ssig;
Analysis:
    Name = Test1;
    Mode = Calibration;
Title:
2 Dim Graded Response Model with fixed Factor Loading Assignment
Comments:
all items have assigned on one of the factors
Estimation:
    Method = BAEM;
    E-Step = 500, 1e-005;
    SE = S-EM;
    M-Step = 50, 1e-006;
    Quadrature = 49, 6;
    SEM = 0.001;
    SS = 1e-005;
Scoring:
    Mean = 0;
    SD = 1;
Miscellaneous:
    Decimal = 2;
```

```
    Processor = 1;
    Print M2, CTLD, Loadings, P-Nums, Diagnostic;
    Min Exp = 1;
Groups:
Group :
    Dimension = 2;
    Items = all_Q1, all_Q2, all_Q9v3, all_Q3, all_Q4, all_Q5,
    all_Q6, all_Q7, all_Q8, all_Q10, all_Q11;
    Codes(all_Q1) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q2) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q9v3) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q3) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q4) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q5) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q6) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q7) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q8) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q10) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q11) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Model(all_Q1) = Graded;
    Model(all_Q2) = Graded;
    Model(all_Q9v3) = Graded;
    Model(all_Q3) = Graded;
    Model(all_Q4) = Graded;
    Model(all_Q5) = Graded;
    Model(all_Q6) = Graded;
    Model(all_Q7) = Graded;
    Model(all_Q8) = Graded;
    Model(all_Q10) = Graded;
```

```
    Model(all_Q11) = Graded;
    Means = 0.0, 0.0;
    Covariances = 1.0,
                  Free, 1.0;
Constraints:
    (all_Q1, Slope[0]) = 0.0;
    (all_Q2, Slope[1]) = 0.0;
    (all_Q9v3, Slope[0]) = 0.0;
    (all_Q3, Slope[0]) = 0.0;
    (all_Q4, Slope[1]) = 0.0;
    (all_Q5, Slope[0]) = 0.0;
    (all_Q6, Slope[1]) = 0.0;
    (all_Q7, Slope[1]) = 0.0;
    (all_Q8, Slope[1]) = 0.0;
    (all_Q10, Slope[1]) = 0.0;
    (all_Q11, Slope[0]) = 0.0;
```

C.2.3 2Dim GRM-ESEM

2-Dimensional Graded Response Model with Partial Constraint (Exploratory Structural Equation Model)

```
Project:
    Name = 2D_GRM_par;
Data:
    File = .\2D_GRM_par.ssig;
Analysis:
    Name = Test1;
    Mode = Calibration;
Title:
```

```
2 Dim Graded Response Model with Partially fixed Factor
Loading Assignment
Comments:
Items are fixed factor except Q2, Q3, and Q11
Estimation:
    Method = BAEM;
    E-Step = 500, 1e-005;
    SE = S-EM;
    M-Step = 50, 1e-006;
    Quadrature = 49, 6;
    SEM = 0.001;
    SS = 1e-005;
Scoring:
    Mean = 0;
    SD = 1;
Miscellaneous:
    Decimal = 2;
    Processor = 1;
    Print M2, CTLD, Loadings, P-Nums, Diagnostic;
    Min Exp = 1;
Groups:
Group :
    Dimension = 2;
    Items = all_Q1, all_Q2, all_Q9v3, all_Q3, all_Q4, all_Q5,
    all_Q6, all_Q7, all_Q8, all_Q10, all_Q11;
    Codes(all_Q1) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q2) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q9v3) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q3) = 1(0), 2(1), 3(2), 4(3), 5(4);
```

```
    Codes(all_Q4) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q5) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q6) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q7) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q8) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q10) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q11) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Model(all_Q1) = Graded;
    Model(all_Q2) = Graded;
    Model(all_Q9v3) = Graded;
    Model(all_Q3) = Graded;
    Model(all_Q4) = Graded;
    Model(all_Q5) = Graded;
    Model(all_Q6) = Graded;
    Model(all_Q7) = Graded;
    Model(all_Q8) = Graded;
    Model(all_Q10) = Graded;
    Model(all_Q11) = Graded;
    Means = 0.0, 0.0;
    Covariances = 1.0,
                  Free, 1.0;
Constraints:
    (all_Q1, Slope[1]) = 0.0;
    (all_Q9v3, Slope[1]) = 0.0;
    (all_Q4, Slope[0]) = 0.0;
    (all_Q5, Slope[1]) = 0.0;
    (all_Q6, Slope[0]) = 0.0;
    (all_Q7, Slope[0]) = 0.0;
    (all_Q8, Slope[0]) = 0.0;
```

```
    (all_Q10, Slope[0]) = 0.0;
```

2Dim GRM-Explore

2-Dimensional Graded Response Model without Constraint.

```
Project:
    Name = 2D_GRM_no;
Data:
    File = .\2D_GRM_no.ssig;
Analysis:
    Name = Test1;
    Mode = Calibration;
Title:
2 dim graded response model without factor loading assignment
Comments:
all items can associate with both factors
Estimation:
    Method = BAEM;
    E-Step = 500, 1e-005;
    SE = S-EM;
    M-Step = 50, 1e-006;
    Quadrature = 49, 6;
    SEM = 0.001;
    SS = 1e-005;
Scoring:
    Mean = 0;
    SD = 1;
Miscellaneous:
    Decimal = 2;
```

```
    Processor = 1;
    Print M2, CTLD, Loadings, P-Nums, Diagnostic;
    Min Exp = 1;
Groups:
Group :
    Dimension = 2;
    Items = all_Q1, all_Q2, all_Q9v3, all_Q3, all_Q4, all_Q5,
    all_Q6, all_Q7, all_Q8, all_Q10, all_Q11;
    Codes(all_Q1) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q2) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q9v3) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q3) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q4) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q5) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q6) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q7) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q8) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q10) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Codes(all_Q11) = 1(0), 2(1), 3(2), 4(3), 5(4);
    Model(all_Q1) = Graded;
    Model(all_Q2) = Graded;
    Model(all_Q9v3) = Graded;
    Model(all_Q3) = Graded;
    Model(all_Q4) = Graded;
    Model(all_Q5) = Graded;
    Model(all_Q6) = Graded;
    Model(all_Q7) = Graded;
    Model(all_Q8) = Graded;
    Model(all_Q10) = Graded;
```

```
    Model(all_Q11) = Graded;
    Means = 0.0, 0.0;
    Covariances = 1.0,
                  0.0, 1.0;
Constraints:
```