



Improving Health Measures: Evidence From a List Experiment, Cognitive Interviews, and a Vignette Study

Citation

Su, Yanfang. 2015. Improving Health Measures: Evidence From a List Experiment, Cognitive Interviews, and a Vignette Study. Doctoral dissertation, Harvard T.H. Chan School of Public Health.

Permanent link

<http://nrs.harvard.edu/urn-3:HUL.InstRepos:16121144>

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)

**IMPROVING HEALTH MEASURES:
EVIDENCE FROM A LIST EXPERIMENT, COGNITIVE INTERVIEWS, AND A
VIGNETTE STUDY**

YANFANG SU

A Dissertation Submitted to the Faculty of
The Harvard T.H. Chan School of Public Health
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Science
in the Department of Global Health and Population
Harvard University
Boston, Massachusetts.

May, 2015

Improving Health Measures:
Evidence from a List Experiment, Cognitive Interviews, and a Vignette Study

Abstract

Measuring health through surveys is challenging. Participants may respond in a socially favorable but untruthful way, and responses across respondents may be difficult to compare. List experiments and anchoring vignette techniques have been proposed to improve survey measures, and this dissertation applied qualitative and quantitative methods to evaluate either the usefulness or validity of those techniques.

The first paper explored how list-based questions perform compared to direct questions when measuring two behaviors, smoking and intravenous infusion use. The difference-in-differences between the two survey methods and two behaviors was non-significant. List experiments might introduce downward biases rather than alleviate them due to cognitive difficulty in responding.

The second paper applied cognitive interviewing to the design of vignettes among Chinese students with objectively different visual acuities. Ten major problems were identified regarding vignette comprehension, judgment, and responses. Respondents rated vignette character's vision differently from their own, demonstrating a norm of being "strict with oneself and lenient towards others" in an Asian context.

The third paper assessed the validity of three vignette survey methods (i.e., indirect comparison between self and vignettes, direct comparison between self and vignettes, and self-assessment primed by vignettes) in measuring distance vision. Surprisingly, it was found that vignette methods were not improvements over self-assessment, and indirect comparison performed the worst.

These three studies shed light on a series of cognitive difficulties related to advanced survey techniques. Traditional self-assessment may be as useful as the list experiment and more valid than some anchoring vignette techniques in these studies' context.

Key words: self-reports; social desirability bias; interpersonal incomparability; list experiment; anchoring vignettes; survey experiment; cognitive interviews; priming effect; receiver operating characteristic (ROC); the area under the ROC curve; smoking; intravenous infusion use; vision; China

TABLE OF CONTENTS

Paper I	1
Abstract	1
I. Introduction	3
A. Social desirability	3
B. The list experiment	4
C. Intravenous infusion use and smoking in China	6
II. Experimental design	9
A. Hypothesis	9
B. Recruitment, consent and survey procedures	9
C. Survey instruments	10
D. Randomization	14
III. Estimation strategy	14
A. Outcomes and equation form of hypothesis	14
B. Estimating prevalence, difference in prevalence levels and DID	18
IV. Results	19
A. Recruitment and demographic characteristics	19
B. Test of assumptions	20
C. Main results	22
V. Discussion and limitations	24
VI. Conclusion	29
Human subject research ethics	30
Paper II	31
Abstract	31
I. Background	33
II. Materials	36
A. Survey questions	36
B. Assumptions: Vignette equivalence and response consistency	38
III. Methods	38
A. Sample	38
B. Procedures	39

C.	Cognitive interviews	40
D.	Objective measure of visual acuity	41
IV.	Data analysis and findings	42
A.	Characteristics of the participants	42
B.	Comprehension I: Hypothetical person.....	42
C.	Comprehension II: Vignette equivalence.....	44
D.	Judgment I: Reference point	47
E.	Judgment II: Leading frame that implies a “right answer”	49
F.	Response I: Challenges to response consistency	49
G.	Response II: Comparison of vignettes	52
H.	Response III: Comparison of Vignette methods	54
I.	Designing vignettes.....	57
V.	Discussion and limitations	59
VI.	Conclusion and implications.....	62
	Competing interests	64
	Human subject research ethics.....	64
	Paper III	65
	Abstract.....	65
I.	Introduction.....	68
II.	Indirect comparison in anchoring vignettes	72
A.	Function	72
B.	Assumptions of indirect comparison.....	74
C.	Improving anchoring vignettes	75
III.	Experimental methods.....	77
A.	Research question and hypotheses	78
B.	Procedures.....	78
C.	Measures	81
IV.	Estimation strategy.....	84
A.	Outcomes	84
B.	The areas under the ROC curves.....	88
V.	Results.....	91
A.	Descriptive characteristics	91

B. Main results.....	91
VI. Discussion and limitations	96
VII. Conclusion and implications.....	100
Competing interests	102
Human subject research ethics.....	102
Clinical trial registration	102
Appendices.....	103
Appendix 1.1. Paper I: Design of survey instruments, pilot, and power calculation.....	103
A. Cognitive Interviewing	105
B. Pilot and Power Calculation.....	110
Appendix 1.2. Paper I: Construction of dependent variables	111
A. Indirect estimates of prevalence: Smoking and intravenous infusion use	111
B. Difference of prevalence levels: Smoking and intravenous infusion use	112
C. Difference-in-differences	114
Appendix 2.1. Paper II: Question appraisal to identify potential source of error in survey	115
Appendix 2.2. Paper II: Cognitive interview protocol.....	118
Condition A (Non-comparative judgment)	118
Condition B (Comparative judgment).....	122
Appendix 2.3. Paper II: Final version of questionnaire	125
Condition A (Non-comparative judgment)	125
Condition B (Comparative judgment).....	126
Appendix 3.1. Paper III: Pilot and power calculation.....	127
Appendix 3.2. Paper III: Simplified Snellen chart.....	128
Bibliography	129

FIGURES WITH CAPTIONS

Figure 3.1.	Response scales and vignettes to calibrate self-reports.....	73
Figure 3.2.	Sample of ROC curve	90
Figure 3.3.	ROC curves, by survey technique.....	94

TABLES WITH CAPTIONS

Table 1.1.	List-based and direct responses, by randomized group.....	14
Table 1.2.	Outcome measures, indicators and mathematical equations	16
Table 1.3.	Demographic characteristics	20
Table 1.4.	Pre-intervention demographic characteristics, by randomized group	20
Table 1.5.	The proportion of students by response value.....	21
Table 1.6.	Prevalence, difference in two prevalence levels, and DID	23
Table 1.7.	Intravenous infusion use among students	104
Table 2.1.	Two versions of survey questionnaires	37
Table 2.2.	Sample distribution	39
Table 2.3.	Sequence of research activities	40
Table 2.4.	Demographic characteristics	42
Table 2.5.	Vignette comprehension	43
Table 2.6.	Difficulty in responding to vignettes, by objective vision	53
Table 2.7.	Difficulty in responding to question formats, by objective vision	55
Table 3.1.	Assumptions of vignette techniques.....	76
Table 3.2.	Randomization strategy.....	80
Table 3.3.	Randomized survey questionnaires.....	82
Table 3.5.	Demographic characteristics	91
Table 3.6.	Pre-intervention demographic characteristics, by randomized group	92
Table 3.7.	Ranking inconsistency in direct comparison and indirect comparison	92
Table 3.8.	Direct comparison: Distribution of responses.....	93
Table 3.9.	Areas under the ROC curves, by survey technique.....	94
Table 3.10.	Pearson Chi-square tests of the areas under the ROC curves	95
Table 3.11.	Improvement from priming effect and using self-assessment	96

ACKNOWLEDGEMENTS

I am greatly appreciative to my research committee - Professors William Hsiao (chair), Joshua Salomon, and Margaret McConnell. They have advised me closely in the course of dissertation development and I deeply appreciate their mentorship, time, and efforts over the past few years. I would also like to thank my oral committee member, Professor Gary King, for his advice at important steps of the research process.

For Paper 1, I am thankful to Professor Zhongliang Zhou who helped with study implementation and data entry. I am thankful to students in Xi'an Jiaotong University Medical School, China, who participated in the study. I have benefited from helpful comments by Julian Jamison, Adam Glynn, Chase Harrison, Jesse Heitner, and Jennifer Pan, as well as participants in the Doctoral Research Seminar at Harvard and the North East Universities Development Consortium. I am indebted to the B&P Foundation (Hong Kong) for financial support.

For Paper 2, I am thankful to the participating school, Jingning Middle School, China, and to Fangfang Xing, Huangzhong Ji, and Jiafeng Wu for logistics support. I would like to especially thank Dr. Gordon Willis for his advice and expertise in cognitive interviewing. I am thankful to students who devoted time and effort for the interviews. I have benefited from careful editing by Yang Cai, Xiaoyu Pan, and Anne Watt. I am thankful for comments by Carol Cosenza, Chase Harrison, and by participants of the Harvard-MIT-BU Chinese Politics Workshop. I am indebted to the B&P Foundation (Hong Kong) for financial support.

For Paper 3, I would like to thank the Jingning County Government and the Bureau of Health for establishing a project coordination office. In addition, I thank Fangfang Xing, Linyu Su, Qian Yang, Shishen Chen, Qiuying Xia, Fang Ye, Rongfen Mao, and Jiapeng Wu for their help with study implementation. I am thankful to the participating students from five schools in Jingning

County, China. I am thankful to Huangzhong Ji, Jianxin Zhou, and Yongqi Pan for their assistance in data entry. I am thankful for comments by Jesse Heitner, Chase Harrison, and by participants of the Harvard-MIT-BU Chinese Politics Workshop. I am also indebted to Harvard T.H. Chan School of Public Health for financial support.

I am deeply thankful for careful editing of all three papers by my dear friend Ben Campbell. I am also thankful to the outside reader, Professor Arie Kapteyn, for helpful comments. I have benefited from the assistance of librarians Scott Lapinski and Carol Mita. I am thankful to Dr. John Watt, Professor Richard Levins and Professor Yuanli Liu for conversations on public health in China from historical, ecological, and practical perspectives.

I would like to especially thank Barbara Heil, Laura Ruggiero, and Anne Watt, for ongoing dissertation support. I am also grateful to my friends Jesse Heitner, Victoria Fan, Yi-Sheng Chao, Mingqiang Li, and Changzheng Yuan for peer conversations. I am indebted for the Departmental Scholarships at the Harvard T.H. Chan School of Public Health and for the Desmond and Whitney Shum Fellowship at the Fairbank Center for Chinese Studies at Harvard.

I wish to dedicate this dissertation to my family, who have supported me at every step of my education and career. I especially dedicate this dissertation to my parents, Yihong Su and Mulian Wu, my parents-in-law, Shaochuan Wu and Guoxiang Lin, and my dear sister, Yanfen Su. Last but not least, I dedicate my years of doctoral work at Harvard to my husband, Zhuoqing Wu, who has made it possible, and to our beloved daughter, Sophie Wu.

Yanfang Su

April 2015

Paper I

Direct Questioning or List-based Questioning:

Evidence from a Survey Experiment on Intravenous Infusion Use and Smoking in China

Abstract

Background Measuring health through surveys is challenging because participants may respond in a socially favorable but untruthful way. To overcome this social desirability bias, attempts have been made to measure human behaviors through complex indirect questioning methods, such as the list experiment. This study compared a list experiment questioning strategy to the standard direct questioning method for two behaviors, intravenous infusion use and smoking. It was expected that intravenous infusion use would be perceived as being socially desirable or neutral and smoking as being socially undesirable by students. The hypothesis was that indirect questioning would increase the reporting of smoking compared to direct questioning, and that the gap between indirect questioning and direct questioning would be significantly larger for smoking than for intravenous infusion use.

Methods A survey experiment was designed to measure the prevalence of intravenous infusion use and smoking among medical students in China by both direct and list-based questions. In a two-by-two design, two groups were asked to respond to a list-based control question, followed by direct questions on either smoking or intravenous infusion use. The second two groups responded to list-based questions about smoking or intravenous infusion use, followed by a direct placebo question.

Results Data were collected from 1,439 medical students. The estimated prevalence of smoking from indirect and direct questions was 4% and 8%, respectively, but the 4% negative difference was non-significant. The estimated prevalence of intravenous infusion use from indirect and direct questions was 43% and 52%, respectively, but the 9% negative difference was non-significant. The difference in differences was 5%, which was not significantly different from zero.

Conclusions The list experiment yielded lower point estimates of prevalence than direct questioning for smoking as well as intravenous infusion use, but the findings were non-significant. These findings contradict the assumption that smoking should show higher estimates using an indirect question compared to a direct question if smoking was socially undesirable. List experiments might introduce downward biases rather than alleviate them due to cognitive difficulty in responding. List experiments might not be more suitable than the anonymous self-administered direct method for measuring health behaviors.

Key words: self-reports; social desirability bias; list experiment; smoking; intravenous infusion use; China

I. Introduction

The tendency of respondents to answer survey questions in a manner that is viewed favorably by others suggests a social desirability bias (Sudman, Bradburn et al. 1996; King and Bruner 2000). Direct questioning might be more prone to social desirability bias than indirect questioning through a list experiment (i.e., item count technique). In list experiments, individual responses about sensitive topics are not collected; instead, a respondent only indicates the total number of statements that apply in the list. This study was designed to compare self-administered direct questioning and list-based questioning when measuring the prevalence of intravenous infusion use and smoking. The intent of the study was to examine the extent to which list experiments elicit distinctive prevalence levels of two health behaviors, which hypothetically have differing degrees of social desirability.

A. Social desirability

In the psychology literature, social desirability was first interpreted as a personality characteristic and the measurement of social desirability has evolved over time. Crowne and Marlowe (1960) developed a test to measure social desirability as a personality trait. The Marlowe-Crowne Social Desirability Scale consisted of 33 true/false items and generated a score indicating a high or low tendency of a person to provide socially desirable responses (Crowne and Marlowe 1960). Then, in 1991, Paulhus developed another method, the Balanced Inventory of Desirable Responding, a questionnaire designed to measure two forms of socially desirable responses (Paulhus 1988). This 40-item instrument provided separate subscales for “impression management,” when there was a tendency not to be honest, and “self-deceptive enhancement,”

when there was a tendency to give honest but inflated descriptions (Paulhus 1988). In self-evaluation on social desirability scales in China, it has been shown that for college students, the need to enhance one's image might take precedence over the need to be honest (Liu, Xiao et al. 2003). Rather than reflecting a constant personality trait, social desirability varies by the nature of the topic.

B. The list experiment

In efforts to address social desirability bias, complex indirect survey techniques have been developed (Raghavarao and Federer 1979; Nederhof 1985; Fisher 1993). One of the most popular indirect survey methods is known as the item count technique (Droitcour, Caspar et al. 1991; Dalton, Wimbush et al. 1994) or the list experiment (Kuklinski, Cobb et al. 1997). In the list-based question, respondents indicate the total number of statements that apply to him or her in the list. It has been argued that an aggregated response to a list of statements is less sensitive than individual responses to a single question. When a respondent is asked how many statements in a list apply to them, he or she is more likely to reveal an accurate answer, even if the list contains sensitive statements. Conducted properly, the list experiment may be a more suitable tool than direct questioning when measuring sensitive health behaviors.

The design of a list experiment involves multiple parts: a key statement (i.e., the statement mentioning sensitive behavior), several non-key statements, and a placebo statement (Please refer to Table 1.8 in Appendix 1.1 for an example of a list experiment). In a treated list, a key

statement is accompanied by several non-key statements. A control list is identical to the treated list, except the key statement is replaced by a placebo statement (i.e., a statement that has been determined to be highly unlikely to be true among the target population). By examining the differences in responses between the randomized treated and control lists, researchers can estimate the prevalence of the sensitive behavior.

Researchers hypothesize that the indirect survey techniques reduce social desirability bias by protecting the privacy of respondents (De Jong, Pieters et al. 2010), and there is some evidence corroborating this. For example, studies have shown that list experiments can reduce over-reporting of positively perceived behaviors such as church attendance (Presser and Stinson 1998), voter turnout (Belli, Traugott et al. 1999; Burden 2000; Holbrook and Krosnick 2010; Comşa and Postelnicu 2013) and “sense of purpose” in work motivation (Antin and Shaw 2012). List experiments can reduce under-reporting of undesirable behaviors such as abortion (Jones and Forrest 1992), drug use (Falck, Siegal et al. 1992; McNagny and Parker 1992; Fendrich and Vaughn 1994; McElrath, Dunham et al. 1995), sexual risk behavior (LaBrie and Earleywine 2000), anti-gay sentiment (Coffman, Coffman et al. 2013) and “killing time” as a work motivation (Antin and Shaw 2012).

However, in other studies using list experiments, results have been mixed. For instance, some studies found that drug use was more detectable in a list experiment than in direct questioning (Falck, Siegal et al. 1992; McNagny and Parker 1992; Fendrich and Vaughn 1994; McElrath,

Dunham et al. 1995), while another study found that the behavior was equally detectable by both a list experiment and direct questioning (Droitcour, Caspar et al. 1991). Given the mixed results in the research to date, more evidence is needed to address the usefulness of the list experiment (Tsuchiya, Hirai et al. 2007).

C. Intravenous infusion use and smoking in China

Intravenous infusion use^{*} was chosen as a target behavior because of its widespread and inappropriate use in China, as described below. Smoking was chosen as the secondary target behavior of this study because, as described below, it has been perceived as socially undesirable, allowing it to be an anchor for comparative analysis for intravenous infusion use. Specifically, this study assumes negative social desirability bias in self-reporting of smoking as well as positive or indistinguishable social desirability bias in self-reporting of intravenous infusion use. To our knowledge, the social desirability biases of these two behaviors have not yet been measured in China, and this study aims to address this gap.

It is likely that intravenous infusion use is socially desirable or neutral among the young population in China. Given that intravenous infusion is mainly used for administration of antibiotics in China (Currie, Lin et al. 2011), few microbiological tests were conducted prior to antibiotic prescribing (Hu, Liu et al. 2003). Studies have shown that doctors are incentivized to offer intravenous infusions because they are more profitable than oral medicines (Sun, Jackson et

^{*} An intravenous infusion is the infusion of liquid substances directly into a vein from a drip chamber.

al. 2009). For example, one study showed that although health workers knew about the use of oral rehydration solution for diarrhea, intravenous infusions were frequently used to treat mild dehydration (Hesketh and Zhu 1997). Besides delivery of antibiotics to combat illness, intravenous infusions have also become more common for healthy students in highly competitive academic settings.

The World Health Organization (WHO) recommends that intravenous infusion be used only for managing extreme illness and for situations in which fluids cannot be taken orally among school children because of the potential risk and harm of intravenous infusion for children. Specifically, the intravenous route is recommended only for management of severe dehydration, septic shock, delivering intravenous antibiotics, and for when oral fluids are contraindicated (such as those with perforation of the intestine or other surgical abdominal problems) (WHO 2005; WHO 2013a). In countries with high compliance to the WHO recommendations, only very poor health status or severe situations lead to intravenous infusion use. While the immediate effectiveness of intravenous infusions compared to oral medication is recognized in China, the safety concerns proclaimed by the WHO have not been widely publicized.

Self-reported smoking is likely to be subject to social desirability bias when solicited via survey. Since 1950, more than 70,000 scientific papers have isolated the causal relationship of smoking and a wide variety of ailments, constituting the largest and best documented body of literature linking any behavior to disease in humans (CDC 1994). The WHO warned about the dangers of

tobacco in a major report on global tobacco control in 2011 (WHO 2011). Although the smoking prevalence rate was decreasing in China (MOH 2006; Li, Hsia et al. 2011), China is the largest tobacco consumer in the world, with 301 million current smokers within the country (Li, Hsia et al. 2011). Further, children's positive attitude towards smoking was associated with tobacco advertisements (Lam, Chung et al. 1997). The WHO has urged bans on tobacco advertising, promotion and sponsorship (WHO 2013b). Given that anti-tobacco educational campaigns have been conducted in China for more than a decade, the awareness of harms from smoking has increased (Huang, Thrasher et al. 2014). Therefore, it is expected that there will be a greater level of reporting of smoking from a list experiment than that from direct questioning.

In this study, based on the theory of social desirability bias, we investigate whether a larger difference in measured prevalence exists between direct and list-based questions for smoking than for intravenous infusion use. The underlying rationale is that participants might face a conflict between the desire to reveal the correct answer and the desire to give the socially favorable response when reporting health behaviors. Additionally, given the cost of intravenous infusion use (Zhang, Eggleston et al. 2006; Xiao, Hou et al. 2010; Zeng and Cai 2011) and smoking (MOH 2006) to the health system in China, it is important to understand the prevalence and the social desirability of these behaviors.

II. Experimental design

A survey experiment was conducted, in which both direct questions and list-based questions were designed. All participants were randomized into four groups at the individual level to test the relative social desirability bias between intravenous infusion use and smoking.

A. Hypothesis

The null hypothesis was that indirect questioning would yield an equal difference in estimated prevalence levels from direct questioning for both behaviors, smoking and intravenous infusion use. The ex-ante alternative hypothesis was that a larger positive measured difference of prevalence would exist between list-based and direct questions for smoking than for intravenous infusion use. Specifically, using a behavior assumed to have non-negative social desirability bias (i.e., intravenous infusion use) as a comparison, the ex-ante expectation was that indirect questioning would yield a significantly higher estimated prevalence than direct questioning for a behavior with a negative social desirability bias (i.e., smoking). In the case that the alternative hypothesis was accepted, it would be inferred that intravenous infusion use was socially more acceptable than smoking.

B. Recruitment, consent and survey procedures

The experiment was carried out among 1,439 students in Xi'an Jiaotong University Medical School, Shanxi Province in northwestern China, in May and June, 2014. Only adult students

aged 18 years or older were recruited for this study. The recruitment of students occurred in a classroom setting.

Each student responded to a short survey that was self-administered. In the survey, the following information was collected: program (undergraduate or not), the year started the program, hometown province, rural/urban, age, gender, father's educational level and mother's educational level, intravenous infusion use, smoking, and visit of Taiwan.

C. Survey instruments

The list experiment in this study was designed according to suggested best practices, such as using in-depth interviewing (Droitcour, Caspar et al. 1991), determining the optimal number of non-key statements (Corstange 2009; Comşa and Postelnicu 2013; Glynn 2013), and testing the underlying assumptions (Holbrook, Green et al. 2003; Martinez 2003; Blair and Imai 2012).

First, because the design of a list experiment can be improved by cognitive interviewing (Droitcour, Caspar et al. 1991), two rounds of cognitive interviews were applied during the development stages of the list experiment (Appendix 1.1).

Second, determining the number of non-key statements is an important component in the design of a list experiment. A simulation study showed that, as the number of non-key statements increased, the standard error of the point estimate of sensitive behavior increased (Corstange

2009). Additionally, too many statements might make it too cognitively difficult to respond. However, if the total number of non-key statements increases or if the non-key statements are negatively correlated, it is less likely that the respondent affirms or denies all non-key statements (Glynn 2013), thereby making it less likely that the respondent is forced to inadvertently reveal the answer to the key statement by having all “yes” or all “no” answers. Researchers have suggested that using four or less non-key statements in a list experiment is ideal (Tsuchiya, Hirai et al. 2007; Comşa and Postelnicu 2013).

Taking these points into account, twelve statements were designed in the phase of pretesting and four statements were selected from the pool to form two pairs of negatively correlated statements (Appendix 1.1). The following represents the control list that was used in the study:

- I performed better in math than Chinese in Grade 12. (Non-key statement 1)
- I fell asleep during class at least once in Grade 12. (Non-key statement 2)
- I visited Gaoxiong, a city in Southern Taiwan, in Grade 12. (Placebo statement)
- I practiced calligraphy in Grade 12. (Non-key statement 3)
- I spent time reading novels in Grade 12. (Non-key statement 4)

With the same four non-key statements, the placebo statement in the control list was replaced with the key statement in the treated list. The following were two key statements in the treated lists.

- I smoked at least one cigarette in Grade 12. (Key statement about Smoking)

- I had an intravenous infusion, *commonly known as ‘dripping infusion,’* in Grade 12. (Key statement about intravenous infusion use)

The survey question was, “How many of the following statements were true for you in Grade 12? (Please indicate the total number but not which ones in particular.)” The students were instructed to write down the number with the explanation that, “The answer ranges from 0 to 5. Please fill 0 if none of the statements apply to you. Please fill 5 if all statements apply to you.”

Third, there are important underlying assumptions to satisfy in the list experiment (Holbrook, Green et al. 2003; Martinez 2003; Blair and Imai 2012). Violation of these assumptions might introduce bias and yield little benefit in using a list experiment in improving the measurement of behaviors. Potential biases in the list experiment are addressed in this study and these biases are important to consider in the interpretation of results.

There are four important assumptions to consider. The first assumption (Assumption I) is a balance in randomization. The pre-intervention characteristics of two randomized groups should be the same, which can be demonstrated by showing that the demographic characteristics in the treated and the control groups are not significantly different. The second assumption (Assumption II) is that the response to non-key statements is independent from the presence of the key statement. In the case that the presence of the key statement induces the student to over-report or under-report the behaviors in the non-key statements, the imputed estimate of the target

behavior would be biased. The independence between the non-key statements and the key statement also ensures the efficiency of the estimate from a list experiment, by eliminating the covariance between the non-key statements and the key statement. The third assumption (Assumption III) is that there is a truthful response to the key statement in the list experiment. It is assumed that, for socially desirable or undesirable behaviors, the student responds to the key statement truthfully, even though they might under-report or over-report in direct questioning. However, when all of the non-key statements apply or do not apply to the student, the protection of privacy in the list experiment vanishes. If the student answers ‘no’ to all non-key statements, the student might over-report socially desirable behaviors (floor effect); when the student answers ‘yes’ to all non-key statements, the student might under-report the sensitive behaviors (ceiling effect). The fourth assumption (Assumption IV) is that there is no design effect. It is required that no difference in cognitive difficulty exists in responding to the treated list and the control list. If the key statement adds significant cognitive difficulty to counting up the total number of statements, this assumption would be violated.

For direct questioning, intravenous infusion use was estimated by the following question: “Did you have an intravenous infusion, commonly known as ‘dripping infusion,’ in Grade 12?”

Smoking was estimated by the following question: “Did you ever smoke at least one cigarette in Grade 12?” The placebo question was, “Did you visit Gaoxiong, a city in Southern Taiwan, in Grade 12?” Those three questions all generated binary responses of ‘yes’ or ‘no.’

D. Randomization

All students were randomized into four groups at the individual level in a two-by-two scheme with equal probability (Table 1.1). The list experiment about smoking consisted of a control list and a treatment list for smoking; the same design was used for intravenous infusion use.

Therefore, 25% of the students were randomly assigned to each of the four groups in Table 1.1.

Each student first responded to the list-based question, followed by the direct question (Table 1.1).

Table 1.1. List-based and direct responses, by randomized group

Randomized groups (1)	List-based responses (List) (2)	Direct responses (Direct) (3)
Smoking _{DirectQ}	List _{Control1}	Direct _{Smoking}
IV _{DirectQ}	List _{Control2}	Direct _{IV}
Smoking _{List}	List _{Smoking}	Direct _{Placebo1}
IV _{List}	List _{IV}	Direct _{Placebo2}

Note: In column (1), Smoking_{DirectQ} and IV_{DirectQ} represent that smoking or intravenous infusion use was asked through direct questioning, respectively; Smoking_{List} and IV_{List} represent that smoking statement or the statement about intravenous infusion use was buried in the list, respectively.

The contents of list-based and direct responses vary by randomized groups, specified in column (2) and column (3), respectively. In column (2), both List_{Control1} and List_{Control2} are responses to the same control list, consisting of four non-key statements and a placebo statement. List_{Smoking} is responses to a treated list, consisting of four non-key statements and a statement about smoking. List_{IV} is responses to the other treated list, consisting of four non-key statements and a statement about intravenous infusion use.

In column (3), Direct_{Smoking} and Direct_{IV} are responses to the direct question about smoking and intravenous infusion use, respectively. Direct_{Placebo1} and Direct_{Placebo2} are responses to the same placebo question about visiting a city in Taiwan.

III. Estimation strategy

A. Outcomes and equation form of hypothesis

The experiment was designed to ultimately estimate the prevalence levels of health behaviors, the difference of prevalence levels between direct questioning and indirect questioning, and

difference-in-differences (DID) (Table 1.2). Accordingly, Table 1.2 shows the outcome measures, the indicators, and the mathematical equations.

Table 1.2. Outcome measures, indicators and mathematical equations

Outcome measure	Indicator	Mathematical equation
Prevalence of smoking	Prevalence _{IndirectSmoking}	mean (List _{Smoking}) - mean (List _{ControlPooled}) + mean (Direct _{PlaceboPooled})
	Prevalence _{Directsmoking}	mean (Direct _{Smoking})
Difference #1	Difference _{Smoking}	Prevalence _{IndirectSmoking} - Prevalence _{Directsmoking} = [mean (List _{Smoking}) - mean (List _{ControlPooled}) + mean (Direct _{PlaceboPooled})] - mean (Direct _{Smoking})
Prevalence of IV use	Prevalence _{IndirectIV}	mean (List _{IV}) - mean (List _{ControlPooled}) + mean (Direct _{PlaceboPooled})
	Prevalence _{DirectIV}	mean (Direct _{IV})
Difference #2	Difference _{IV}	Prevalence _{IndirectIV} - Prevalence _{DirectIV} = [mean (List _{IV}) - mean (List _{ControlPooled}) + mean (Direct _{PlaceboPooled})] - mean (Direct _{IV})
Difference #3	DID	Difference _{IV} - Difference _{Smoking} = (Prevalence _{IndirectSmoking} - Prevalence _{Directsmoking}) - (Prevalence _{IndirectIV} - Prevalence _{DirectIV}) = [mean (List _{Smoking}) - mean (Direct _{Smoking})] - [mean (List _{IV}) - mean (Direct _{IV})]

Note: The responses to the control list in two randomized groups are pooled by taking the average of the responses. $List_{ControlPooled} = \frac{List_{Control1} + List_{Control2}}{2}$

The responses to the placebo direct question in the randomized groups are pooled by taking the average of the responses.

$Direct_{PlaceboPooled} = \frac{Direct_{Placebo1} + Direct_{Placebo2}}{2}$

The estimated prevalence of “yes” responses to the key statement can be imputed by subtracting the mean of control list responses from the mean of the treated list responses and then adding the mean of the placebo question responses, as shown in Table 1.2. The following is a simple example how to calculate the prevalence of intravenous infusion use from a list experiment, with a single control list and a single placebo direct question.

$$\begin{aligned}
 & \text{Prevalence}_{\text{IndirectIV}} \\
 &= \text{mean}(\text{List}_{\text{IV}}) - \text{mean}(\text{List}_{\text{Control}}) + \text{mean}(\text{Direct}_{\text{Placebo}}) \\
 &= \text{mean}(\text{non-key statements} + \text{IV statement}) - \text{mean}(\text{non-key statements} + \text{Placebo statement}) + \text{mean}(\text{Direct}_{\text{Placebo}}) \\
 &= \text{mean}(\text{IV statement}) - \text{mean}(\text{Placebo statement}) + \text{mean}(\text{Direct}_{\text{Placebo}}) \\
 &= \text{mean}(\text{IV statement})
 \end{aligned}$$

Similar logic can be applied to estimate other indicators that use data from the list experiment in Table 1.2.

Social desirability bias was measured by the discrepancy between the means of list-based and direct responses, presented as $\text{Difference}_{\text{IV}}$ and $\text{Difference}_{\text{Smoking}}$ (Table 1.2).

In equation form, the null hypothesis was:

$$(\text{Prevalence}_{\text{IndirectSmoking}} - \text{Prevalence}_{\text{Directsmoking}}) - (\text{Prevalence}_{\text{IndirectIV}} - \text{Prevalence}_{\text{DirectIV}}) = 0$$

The alternative hypothesis was:

$$(\text{Prevalence}_{\text{IndirectSmoking}} - \text{Prevalence}_{\text{Directsmoking}}) - (\text{Prevalence}_{\text{IndirectIV}} - \text{Prevalence}_{\text{DirectIV}}) > 0$$

Further, for smoking, it was expected that prevalence would be higher for indirect questioning than for direct questioning, and therefore, $\text{Prevalence}_{\text{IndirectSmoking}} > \text{Prevalence}_{\text{Directsmoking}}$. For intravenous infusion use, it was expected that there would be similar or less prevalence found from indirect questioning than from direct questioning, and therefore,

$$\text{Prevalence}_{\text{IndirectIV}} \leq \text{Prevalence}_{\text{DirectIV}}.$$

B. Estimating prevalence, difference in prevalence levels and DID

Different estimation methods have been used for list experiments (Tsuchiya 2005; Blair and Imai 2012), and because the study design was crafted to provide insight into several important measurement questions, the following methods were used to respond to specific needs in this study. Both least square estimations (LSE) and maximum likelihood estimations (MLE) were applied in data analysis of the list experiment, prevalence differences, and DID, for which the following regression specification was used.

$$Y_i = \beta_0 + \beta_1 \text{IV}_{\text{DirectQ}} + \beta_2 \text{Smoking}_{\text{List}} + \beta_3 \text{IV}_{\text{List}} + \varepsilon_i$$

Y_i indicated a dependent variable, in which i represented the individual student. The dependent variables to estimate the prevalence, the difference of prevalences, and DID are presented in Appendix 1.2. $\text{IV}_{\text{DirectQ}}$, $\text{Smoking}_{\text{List}}$, and IV_{List} were dummy variables for participants who were assigned to the group for which the list included the placebo statement, smoking statement, and statement about intravenous infusion use, respectively. The dummy variables, $\text{DirectQ}_{\text{IV}}$, $\text{List}_{\text{smoking}}$, and List_{IV} took the value ‘1’ if a student got that version of the survey and ‘0’ otherwise. The coefficients, β_1 , β_2 and β_3 , were the discrete difference of Y_i due to the variation of each dummy variable, respectively. The reference group was $\text{Smoking}_{\text{DirectQ}}$, for whom smoking was asked in the direct question. The mean of the dependent variable for the reference group was captured by β_0 .

IV. Results

A. Recruitment and demographic characteristics

Totally, 1,489 students in Xi'an Jiaotong University Medical School were defined as the study population and invited to participate in the study in May and June, 2014. Finally, 1,439 students were recruited, with a participation rate at 97%. Among the recruited students, 1,369 students responded to the survey, with a response rate at 95%. Among the 1,369 students, 5 students that self-reported an age of 17 years old (though they claimed they were 18 years old or older in consent) were excluded in analysis. The students who enrolled for pretesting and the pilot were invited for the large-scale survey as well, due to administrative difficulty in excluding them from the anonymous survey. At the end of the survey, all students were asked, "Did you participate in this survey between November, 2013 and March, 2014," to distinguish the previous participants. Therefore, 59 students that recalled that they responded to the survey in the pretesting and pilot stages were also excluded.

Finally, 1,305 students were included in data analysis and the demographic characteristics are presented in Table 1.3.

Table 1.3. Demographic characteristics

Variable	Obs	Mean
Age	1292	20.6
% of male students	1299	38%
% from rural China	1287	44%
% of hometown in Shanxi	1295	57%
Father edu <12yrs	1302	62%
Mother edu <12yrs	1302	71%
% undergraduates	1304	96%
% freshmen	1304	30%

B. Test of assumptions

First, there was no significant difference among the randomized groups (Assumption I).

Randomization was balanced in terms of age, rural residents, hometown location, parental education, sex ratio, the percentage of undergraduates and the percentage of freshmen (Table 1.4).

Table 1.4. Pre-intervention demographic characteristics, by randomized group

	DirectQ _{smoking}	List _{smoking}	DirectQ _{IV}	List _{IV}	Prob > F
Mean age	21	20	21	21	>0.05
% of students from rural	43%	45%	43%	40%	>0.05
% of hometown in Shanxi	57%	62%	58%	53%	>0.05
Father's edu < 8 yrs	60%	62%	66%	58%	>0.05
Mother's edu < 8 yrs	70%	70%	74%	68%	>0.05
% of male students	38%	41%	38%	37%	>0.05
% of undergraduate students	96%	96%	96%	96%	>0.05
% of freshmen	29%	31%	31%	29%	>0.05
Observations	344	345	338	337	

Second, regarding dependence between the key statement and the non-key statements (Assumption II), in pretesting, Pearson chi-square correlation was conducted between intravenous infusion use and reading novels; no significant correlation was found between the responses to those two behaviors. However, the non-key statements were changed after cognitive interviews; thus, the correlation between behaviors of interest and the other non-key statements remained unknown.

Third, for Assumption III (truthful responses), the null hypothesis was that the percentage of students who answered ‘0’ or ‘5’ in the treated group was greater or equal to that in the control list. The distribution of responses from ‘0’ through ‘5’ is presented in Table 1.5.

Table 1.5. The proportion of students by response value

Response value	Control list			Treated list	
	Control1	Control2	Pooled Control	Smoking in the list	IV in the list
0	2%	3%	3%	3%	2%
1	7%	9%	8%	9%	6%
2	34%	33%	33%	29%	24%
3	42%	42%	42%	42%	37%
4	13%	13%	13%	15%	23%
5	1%	1%	1%	2%	8%
Obs	344	345	689	336	337

T-tests were conducted between the pooled control list and two treated lists and we failed to reject the null of truthful responses. Specifically, in testing for the floor effect, there was no significant difference between the control and the treated groups ($p>0.05$); in testing for the ceiling effect, there was no significant difference between the control and treated list about

smoking ($p > 0.05$) and the percentage of responses with value 5 was significantly higher in the treated list about intravenous infusion use ($p = 0.00$).

In sum, the list experiment met the standards of Assumptions I (balance in randomization) and III (truthful response to the key statement). Assumption II (independence between key statement and the non-key statements) was partially tested. Assumption IV (design effect) could not be adequately assessed in this study.

C. Main results

It was important to first examine the characteristics of list-based and placebo responses.

List-based responses - The mean of list-based responses was 2.98, 2.60, 2.62 and 2.55 for the treated list about intravenous infusion use, the treated list about smoking and two control lists, respectively. The difference in the mean of estimates was 0.07 between two randomized groups responding to the control list, but it was not statistically significant ($p > 0.1$). Thus, the responses to the control list in two randomized groups can be pooled and were pooled for the analysis of the main results.

Placebo responses - For the placebo question, “Did you visit Gaoxiong, a city in Southern Taiwan, in Grade 12,” 1.5% and 2.5% of the students reported they visited Taiwan, in two randomized groups, respectively. The difference between the means of placebo responses was 0.01 but it was not statistically significant ($p > 0.1$). The placebo responses can be pooled and were pooled in estimating the main results.

The estimates from direct questioning, indirect questioning, the difference of prevalence levels between the two survey methods, and difference-in-differences are presented in Table 1.6. There was no missing data in direct questioning and there were only two missing values in indirect questioning, among 1,305 students.

Table 1.6. Prevalence, difference in two prevalence levels, and DID

Indicator	Obs	Models					
		Least Square			Maximum likelihood		
		B(%)	SE	P-value	B(%)	SE	P-value
Prevalence _{IndirectSmoking} ¹	1308	4%	0.07	0.60	4%	0.55	0.95
Prevalence _{Directsmoking} ²	325	8%	0.01	0.00	8%	0.01	0.00
Difference _{Smoking} ³	1308	-4%	0.07	0.55	-4%	0.56	0.94
Prevalence _{IndirectIV} ¹	1308	43%	0.07	0.00	43%	0.79	0.56
Prevalence _{DirectIV} ²	331	52%	0.03	0.00	52%	0.03	0.00
Difference _{IV} ³	1308	-9%	0.08	0.26	-9%	0.84	0.92
DID ⁴	1310	5%	0.09	0.58	5%	0.26	0.85

Note:

¹ Prevalence_{Indirect}: The point estimate of prevalence from indirect questioning is calculated by subtracting the mean of control list responses from the mean of the treated list responses and adding in the mean of the response to the placebo question.

² Prevalence_{Direct}: The point estimate of prevalence in direct questioning is the mean of direct responses.

³ Difference = Prevalence_{Indirect} - Prevalence_{Direct}

⁴ Difference-in-differences (DID) = Difference_{Smoking} - Difference_{IV}

The point estimates were consistent between least square (LS) and maximum likelihood (ML).

However, contradictory to the literature (Blair and Imai 2012; Comşa and Postelnicu 2013; Meng, Pan et al. 2014), the ML estimators yielded larger standard errors in some cases.

Estimated prevalence levels from two methods - The health behaviors were measured both from direct questioning and the list experiment. From the list experiment, the estimated smoking prevalence was 4%, using the pooled control list as the reference group, and the estimator was non-significantly different from zero. From direct questioning, estimated smoking prevalence was 8%, which was significantly different from zero (Table 1.6). From direct questioning, the estimated intravenous infusion use was 52%. From the list experiment, the estimated intravenous infusion use was 43%, using the pooled control list as the reference group. Both point estimates were significantly different from zero (Table 1.6).

Difference of prevalence levels between direct questioning and indirect questioning – The difference in prevalence levels for smoking was 4%, which was negative in sign and non-significant (Table 1.6); the difference in prevalence levels for intravenous infusion use was 9%, which was negative in sign and non-significant.

Difference-in-differences – There was an approximately 5% difference between the two measurements and two behaviors but the estimator was non-significantly different from zero (Table 1.6).

V. Discussion and limitations

A survey experiment was conducted to explore the self-reported prevalence of intravenous infusion use and smoking, with the expectation that indirect questioning would reduce under-

reporting of smoking. The main finding was that the difference-in-differences between direct questioning and indirect questioning for two health behaviors was 5%, which was non-significant. The results failed to reject the null hypothesis that the reporting gap between direct questioning and indirect questioning was the same for intravenous infusion use and smoking among medical students in China.

It was surprising that lower estimates were yielded in the list experiment than direct questioning for smoking. There are several sources of bias that may have influenced this result. First, there is the potential bias resulting from the violation of assumptions. It was estimated that bias was very unlikely to be introduced due to unbalanced randomization (violation of Assumption I), or untruthful responses to the key statement in the list experiment (violation of Assumption III).

The main concern was violation of design effect (Assumption IV). In this study, it was very likely that the measurement error with a downward bias, which occurred in estimating the prevalence levels for both behaviors, was due to counting difficulty. More specifically, it might be sufficiently more difficult to memorize the affirmative answers and add them up in responding to the treated list compared to the control list. Other studies showed that participants' cognitive difficulties in memorizing the affirmative answers and then adding them up introduced measurement errors (Biemer, Jordan et al. 2005; Tsuchiya, Hirai et al. 2007). Other researchers have found similar results. For instance, Droitcour et al. as well as LaBrie and Earleywine applied an unmatched list experiment and they yielded lower estimates in the list experiment than direct questioning for intravenous drug use (Droitcour, Caspar et al. 1991) and college students getting drunk (LaBrie and Earleywine 2000). It was very likely that cognitive difficulty

was greater in responding to the treated list than to the control list because there was an additional statement in the unmatched list experiment.

Another concern was that the response to non-key statements was dependent on the presence of the key statement, leading to a violation of Assumption II. Because this assumption was only partially tested, it is necessary to discuss the likelihood of correlation between the key statement and the non-key statements. The statement of interest was placed in the middle of the list, as the third one out of five. It was possible that the responses to the statements followed by the statement of interest were impacted by the key statement due to order effect (McClendon and O'Brien 1988; Buckley 2008; Lee, Schwarz et al. 2014). It was tested and shown that there was no significant correlation between intravenous infusion use and reading novels in the pretesting phase. However, it was left unknown whether there is a correlation between intravenous infusion use and calligraphy practice as well as between smoking and two non-key statements (i.e., calligraphy practice and reading novels) in the list.

It was also surprising that the smoking prevalence levels estimated through both survey methods in this study were lower than the estimated prevalence in the Global Adult Tobacco Survey (GATS). The GATS sampled from 100 counties/districts in China in 2010, and the estimated prevalence was 18% [95% confidence interval (14.7, 21.6)], among those 15 to 24 years old (Li, Hsia et al. 2011). In this study, the sample was medical students in Xi'an Jiaotong University, with an average age of 20 years old, and smoking prevalence was estimated for the year of 2012. The smoking prevalence was 8% and 4% using the direct and indirect questioning methods, respectively. There are several possible explanations for this discrepancy. First, it is possible that

there were fewer smokers in the medical school in this study than in the nationwide sample. Second, as smoking prevalence declines over time (MOH 2006; Li, Hsia et al. 2011), the estimated prevalence in 2012 could be lower than that in 2010. Third, the cognitive difficulty of responding to the list experiment may have placed a downward bias on the estimate. Relatively, the estimate from direct questioning is closer to the national average estimate than that from the list experiment.

In this study, the results suggest that the list experiment may not be useful in improving the measurement of intravenous infusion use and smoking. Given that there are mixed results from list experiments in the literature, the results from this study belong to the pool of research that has shown no difference between the estimates from a list experiment and direct questioning for the following behaviors: intravenous drug use (Droitcour, Caspar et al. 1991), receptive anal intercourse (Droitcour, Caspar et al. 1991), college students getting drunk (LaBrie and Earleywine 2000), past engagement in counterproductive behaviors (Ahart and Sackett 2004), the prevalence of cocaine use (Biemer, Jordan et al. 2005), giving blood (Tsuchiya, Hirai et al. 2007), and condom use (Jamison and Karlan 2011). Further, counter-intuitive results have been generated from list experiments. For instance, the number of sexual partners was reported higher in direct questioning than in list-based questioning (Jamison and Karlan 2011). In such cases, the ex-anti prior about a specific behavior or the potential bias in a list experiment needs to be examined.

In the field of survey research, perhaps the most effective approach, and the path with minimal levels of social desirability bias, is the use of anonymous, self-administered direct questioning. A

survey on sensitive questions could be self-administered, web-based or telephone-based rather than interviewer-administered so as to avoid interpersonal interaction (Nederhof 1985; Johnson, Hougland et al. 1989). In prior research, when participants were asked to report socially undesirable behavior in a survey free of interviewer presence as opposed to with an interviewer in the room, socially undesirable behavior was reported more frequently when the interviewer was not in the room (Kaminska and Foulsham 2013). This suggests that the likelihood of underreporting a socially undesirable behavior is higher when responding to another person as opposed to when in isolation. Furthermore, a recent report examining online panels by the American Association for Public Opinion Research concluded that, regardless of design, there were higher reports of socially undesirable attitudes and behaviors in self-reported web-based questionnaires than in face-to-face interviews (Baker, Blumberg et al. 2010). In this study, both self-administration of the survey and response without an identifier protected privacy. The low percentage of item non-response suggests that privacy is protected in anonymous self-administration of surveys.

There are several important limitations to this study. First, given that the study sample was medical students, it is difficult to generalize the findings to students in the general population. Second, the prevalence levels of intravenous infusion use and smoking were only measured by surveys rather than objective measurements; therefore, the validity of the survey instruments remains unknown. Third, the surveys were self-administered by participants; therefore, the results of this study cannot be generalized to other survey modes, such as interviewer administration. Fourth, results about cognitive difficulty in responding to the list-based question may not be applicable to other list experiments with less than four non-key statements.

VI. Conclusion

List experiments might not be more suitable than the anonymous self-administered direct method for measuring health behaviors. There was no evidence that list-based questioning yielded greater reports of smoking use when compared to direct questioning. Nor was evidence generated about the level of social desirability for smoking and intravenous infusion use among medical students in China.

The results from this study contradicted the ex-ante assumption that smoking should show a higher estimate of prevalence using list-based questioning than that from direct questioning if smoking was socially undesirable. The surprising finding suggests that the list-based method might introduce downward bias. The bias was plausibly due to the violation of the “no design effect” assumption.

It needs to be acknowledged that it can be a complex task, for participants in a list experiment, to count and memorize the affirmative answers. Even though the number of statements was the same in the control and the treatment groups in the list experiment in this study, the key statement about smoking or intravenous infusion use was more likely to yield an affirmative answer than the placebo statement about visiting Taiwan. Therefore, students in the treated group might experience more counting difficulty in adding up all affirmative answers.

Competing interests

The author declares that she has no competing interests.

Human subject research ethics

This study was reviewed and approved by IRB in Xi'an Jiaotong University Medical School (ID: 2013-231).

Paper II

Designing Vignettes and Question Formats to Measure Distance Vision:

Evidence from Cognitive Interviews among Students in China

Abstract

Background Vignette methods have been proposed as a way to correct comparability problems in self-reported survey measures, but the validity of such methods depends largely on vignette wording and question formatting. This study used cognitive interviewing techniques to evaluate comprehension, judgment, and responses toward vignettes that are used to measure distance vision.

Methods Two vignettes and two vignette approaches were examined through cognitive interviews among 36 students from Grade 7 and Grade 11 in rural China. Interviews were conducted among students with different objective levels of visual ability. The respondents either directly evaluated the vision of the vignette's hypothetical character (i.e., non-comparative judgment) or compared their own vision with the vignette character's vision (i.e., comparative judgment). Data were collected through thinking-aloud, verbal probing, and objective measurement of vision.

Results Ten major problems were identified through cognitive interviewing and the results can be summarized in three categories. First, for vignette comprehension, it was found that the

concept of a hypothetical person was understandable by all participants in Grade 7 and beyond. However, only 50%, 40%, and 25% of the students accurately estimated objective distance of 5 meters, 10 meters, and 20 meters, respectively. Furthermore, ambiguous vignette phrases, such as ‘appear blurry,’ and lengthy wording caused difficulty in comprehension. Second, for vignette judgment, students used either self or a previous vignette as a reference point. Third, for vignette response, due to the Chinese cultural view of being “strict with oneself and lenient towards others,” different standards were applied in self-assessment and the assessment of vignettes by 11% of students. Meanwhile, 79% of students reported that non-comparative judgment was more difficult to answer and the pattern persisted among different visual acuity groups.

Conclusions To our knowledge, this was the first research study applying cognitive interviewing to address the design of vignettes in China. Overall, vignettes about distance vision were understandable for students at the educational levels of grade seven and beyond. It was more difficult to reach an answer and to be certain about the answer in non-comparative judgment than in comparative judgment. Lastly, this study revealed the value of cognitive interviewing in identifying areas for improvement in vignette design within a given cultural context.

Key words: vignette; Self-report; cognitive interviews; non-comparative judgment; comparative judgment; vision; China

I. Background

The anchoring vignette technique has been developed to improve interpersonal comparability in self-reports (King, Murray et al. 2004). Anchoring vignettes have been advanced with tailored statistical methods (Wand, King et al. 2007; Wand 2013), and widely used in the World Health Survey, among other empirical studies (Damacena, Vasconcellos et al. 2005; Chevalier and Fielding 2011; Wada, Kakuma et al. 2011; King, Harper et al. 2012). Importantly, the anchoring vignette technique involves the concept of a hypothetical person as well as the description of a specific scenario in the domain of interest, and it is essential to design the vignette wording and questioning format correctly in order to ensure valid results.

However, vignette design has been largely neglected and it has been suggested that there is a need for further work on the design of vignettes, especially improving vignette descriptions before vignettes are used in practice (Hopkins and King 2010; Kapteyn, Smith et al. 2011a). Specifically, one study suggested that the assumptions underlying anchoring vignette techniques were more likely to hold true if the description of the vignette character's condition was complete and concise (Kapteyn, Smith et al. 2011a). Vignette methods have sometimes been considered infeasible for populations with limited education because participants with lower educational levels are less likely to respond (d'Uva, O'Donnell et al. 2008). The primary purpose of this research study was to use cognitive interviewing to assess how a young population comprehends, judges, and responds to vignettes in the domain of distance vision, and to develop approaches to making vignette methods generally more useful. The secondary purpose was to develop more understandable and sensitive vignettes to measure distance vision in a Chinese context.

Among different health domains, vision was chosen to validate the anchoring vignette technique mainly for two reasons. First, measuring perceptions of visual acuity is an important task. According to one study conducted in China, among children who needed glasses, only 77% of students wore them (Su 2015a). A delay in glasses wearing can be attributed to several factors, including a lack of an objective measure of visual acuity in resource-limited settings and incorrect perceptions of visual acuity. Second, both objective measures and self-assessment are applicable to measuring distance vision (King, Murray et al. 2004). In this study, the systematic cognitive differences among participants with different objective levels of visual ability were explored. This study is in line with other efforts to use cognitive interviewing to improve the measurement of perceived visual acuity (Miller, Mont et al. 2011).

Cognitive interviewing is one way to potentially improve vignette wording and formatting, which is a neglected aspect of survey design (Hopkins and King 2010; Kapteyn, Smith et al. 2011a). Cognitive interviewing is a pretesting method used in questionnaire design that allows researchers to identify problems in question formulation that may prevent them from effectively collecting information (Willis 2005). By administering draft survey questions and then following up with probing questions, a researcher can collect additional verbal information about survey responses and determine whether the question generated the information that the researcher intended (Willis 2005).

Cognitive interviewing techniques have been developed and applied to both government surveys and other social science research in the past three decades (Willis 2005). Since 1988, the practice has been regularly utilized at three United States government statistical agencies - the National Center for Health Statistics, the Census Bureau, and the Bureau of Labor Statistics (Willis 2005). It has been used in pretesting cross-national and cross-cultural surveys (Pasick, Stewart et al. 2001; Nápoles-Springer, Santoyo-Olsson et al. 2006; Nápoles-Springer and Stewart 2006; Willis, Lawrence et al. 2008; Farrall, Priede et al. 2012) and evaluating different sets of questions, such as the Frenchay Activities Index, the Barthel Index, the Simple Lifestyle Indicator Questionnaire, the EuroQoL EQ-5D, and the Personal Resources Questionnaire 2000 (Ploughman, Austin et al. 2010). However, to our knowledge, there has been no research on the use of cognitive interviews to improve vignette design in a Chinese context.

To investigate respondent understanding of vignette questions and to discover potential survey improvements, this study empirically applied cognitive interviewing in vignette development for measuring vision. In cognitive interviews, respondents were encouraged to reveal problems via open-ended discussion rather than being driven to provide closed-ended, coded responses. In particular, comprehension, judgment, and response process were investigated through a standardized interview protocol, according to the four-stage model of the survey response process (Tourangeau, Rips et al. 2000). Understanding these cognitive processes is critical for designing vignettes that can yield informative conclusions on perceptions of health. For purposes of organizing the presentation, the cognitive interviewing reporting framework (CIRF) was adopted in this study (Boeije and Willis 2013).

II. Materials

A. Survey questions

Two initial versions of survey questionnaires were tested in cognitive interviews and there were three questions in each version (Table 3.3). Each questionnaire consisted of prefaces, one self-assessment question and two vignettes. Only vignette questions were phrased in two different ways in two questionnaires. The non-comparative judgment asked them to evaluate their own vision, and the vignette character's vision, directly. Non-comparative judgment is the practice of survey instruments in World Health Surveys, based on which a series of research has been conducted (Damacena, Vasconcellos et al. 2005; Wada, Kakuma et al. 2011; King, Harper et al. 2012). Meanwhile, the comparative judgment asked students to compare their own vision with vignette situations. Comparative judgment has been evaluated by other researchers on the topic of rest/energy (Hopkins and King 2010). Two vignettes were designed to represent scenarios where a hypothetical person had either a distance vision of less than 5 meters or greater than 20 meters. Two hypothetical persons with common Chinese names, Wang Wu, and Zhang San were introduced in vignettes (Table 2.1).

Table 2.1. Two versions of survey questionnaires

Versions	Non-comparative judgment	Comparative judgment
Preface 1	Now, I would like you to think about your own vision <u>without</u> glasses or contact lenses.	
Self-assessment of vision	<p>At the present time, would you say your distance eyesight is:</p> <p>(1) Excellent, (2) Good, (3) Fair, (4) Poor, (5) Very poor.</p>	
Preface 2	When answering the next questions, I want you to think about Wang Wu and Zhang San of your age and gender. Please think about Wang Wu and Zhang San's vision <u>without</u> glasses or contact lenses.	
Vignette questions	<p>[Wang Wu] finds faces to appear blurry at a distance of 5 meters. Would you say [Wang Wu]'s distance eyesight is:</p> <p>(1) Excellent, (2) Good, (3) Fair, (4) Poor, (5) Very poor.</p> <p>[Zhang San] can recognize familiar people's faces and pick out facial expression (e.g., angry, smile) at a distance of 20 meters quite distinctly. Would you say [Zhang San]'s distance eyesight is:</p> <p>(1) Excellent, (2) Good, (3) Fair, (4) Poor, (5) Very poor.</p>	<p>[Wang Wu] finds faces to appear blurry at a distance of 5 meters. Would you say your distance eyesight is:</p> <p>(1) Better than Wang Wu's (2) The same as Wang Wu's (3) Worse than Wang Wu's</p> <p>[Zhang San] can recognize familiar people's faces and pick out facial expression (e.g., angry, smile) at a distance of 20 meters quite distinctly. Would you say your distance eyesight is:</p> <p>(1) Better than Zhang San's (2) The same as Zhang San's (3) Worse than Zhang San's</p>

B. Assumptions: Vignette equivalence and response consistency

Non-comparative judgment is based on two assumptions: vignette equivalence and response consistency (Salomon, Tandon et al. 2004). Vignette equivalence refers to the requirement that underlying domain levels represented in each vignette are understood in approximately the same way by all respondents (Salomon, Tandon et al. 2004). For example, in a study measuring mobility level in Asian countries, interpretations of mobility level, as presented in the vignette, varied significantly among participants across countries (Hirve, Gomez-Olive et al. 2013). This situation represents a violation of vignette equivalence. Response consistency, on the other hand, refers to the requirement that individuals use similar standards in self-assessment and in the evaluation of vignette scenarios (Salomon, Tandon et al. 2004). For instance, one study showed that participants used similar standards for themselves as well as for vignette characters in evaluations of the domain of sleep but not in the domain of pain (Kapteyn, Smith et al. 2011a).

III. Methods

A. Sample

The study was conducted in a middle school in Jingning County, Zhejiang Province, China. A total of 36 interviews were conducted, across two rounds of 18 each. Given that cognitive interviewing normally requires a small sample size, no power calculation was conducted.

The middle schools in China have students between Grade 7 and Grade 12 and the study meant to enroll students to represent largely different educational levels. However, students in Grade 12 were preparing for university entrance exams and they were not invited to attend the study. Any student in Grade 7 or Grade 11 in Jingning Middle School was eligible to participate in the

interview and the age range was determined by that population. In these two grades, students representing a range of objective levels of visual ability were sought for participation in the study.

Children under 18 years old and adult students were recruited for interviews. It was planned to enroll students with particular characteristics of interest (e.g., students who needed glasses but did not have glasses; students who did not need glasses; students who always wore glasses or students who wore glasses sometime), with 25% in each of the four groups. The sample distribution is presented in Table 2.2.

Table 2.2. Sample distribution

Visual status		Good		Poor	
Having glasses		No	No	Yes	
Wearing glasses	Obs	No	No	Always	Not always
First round	18	8	0	5	5
Second round	18	8	0	4	6
Sum	36	16	0	9	11

B. Procedures

The study procedures are summarized in Table 2.3. The designer of this study conducted the interviews and analyzed the data. Before the cognitive interviews, the interviewer conducted a question appraisal (Appendix 2.1) to design the interview protocol. To keep all interviews consistent, one interviewer conducted all interviews. The interviewer conducted interviews of 36 students and each interview was followed by an objective measure of visual acuity. In the first

round, 18 students were interviewed. Findings were utilized to refine the survey questions; and in the next round, an additional 18 students were interviewed to test the refined survey questions. Both qualitative and quantitative analyses were conducted to understand the following aspects: 1) comprehension, 2) judgment, 3) response, and 4) design of vignettes.

Table 2.3. Sequence of research activities

Step 1	Question Appraisal by interviewer First round of study (n=18)
Step 2	(Concurrent interviewing, followed by objective measure of vision)
Step 3	Refinement of questions Second round of study on refined questions (n=18)
Step 4	(Retrospective interviewing, followed by objective measure of distance vision)
Step 5	Finalization of questions

C. Cognitive interviews

In two rounds of interviews, each student only attended one round. The first round of cognitive interviewing involved concurrent probing, in which the subject self-administered a survey question and then the interviewer asked additional questions regarding this specific survey question. The process was repeated three times for the three questions. The second round was retrospective. The student self-administered the entire written survey that consisted of three questions and submitted it to the interviewer; then the interviewer asked additional questions.

In semi-structured interviews, thinking-aloud (Ericsson and Simon 1980) and verbal probing techniques (Forsyth and Lessler 1991) were two critical methods used. Students were allowed and encouraged to think-aloud as the first step. Thinking-aloud is a process of verbalizing one's thought process while arriving at a conclusion (Ericsson and Simon 1980). If this proved difficult, then the interviewer applied probes. Both concurrent and retrospective verbal probes

were applied to facilitate the conversations (See the standardized probes in Appendix 2.2) but there was flexibility in applying predetermined probes or newly developed probes, based on the conversation. The study mainly relied on probing (Appendix 2.2) because it was anticipated to be difficult for both adults and children to think-aloud.

In cognitive interviewing, one randomized version of the vignette questionnaire was presented to the subject: comparative judgment or non-comparative judgment (Table 3.3). A probe on comparative judgment was designed for students who responded to non-comparative judgment and vice versa (Appendix 2.2). By asking the vignette questions in two approaches (i.e., comparative judgment or non-comparative judgment), comparative cognitive difficulty in each approach was tested.

The average interview took about 30 minutes. Most of the interviews were audiotaped. The interviewer took notes while interviewing all of the students. Both the interviewer and interviewees had Chinese as their first language.

D. Objective measure of visual acuity

After interviews, each subject received an objective measure of visual acuity by the “Simplified Snellen Chart” (SSC), consisting of the letter “E” oriented in different directions. The respondent stood 5 meters away from the SSC and indicated the direction of “E”s. The interviewer conducted the test after receiving training from an optometrist in a local setting to ensure safety and accuracy. In the objective measure, distance vision without glasses or contact lenses was

measured even if the subject typically wore visual corrections. The measurement ranged from 4.0 to 5.3 and a larger number indicates better vision.

IV. Data analysis and findings

A. Characteristics of the participants

The characteristics of the study sample are summarized in Table 2.4.

Table 2.4. Demographic characteristics

	Obs	Percent/Mean
Mean age (yrs old)	36	15
% male	36	47%
% in Grade 7	36	50%
% with poor or very poor vision in self-assessment	36	50%
% having glasses	36	56%
% in the first round of interviews	36	50%

In the following sections, we describe findings related to three key challenges in anchoring vignette techniques: comprehension, judgment, and responses to vignettes. The sections are organized into those three categories based on Tourangeau et al.’s psychological model of the survey response (Tourangeau, Rips et al. 2000).

B. Comprehension I: Hypothetical person

All interviewed students in Grade 7 and Grade 11 understood the description of the hypothetical persons (Table 2.5). In the interviews, probes were designed to facilitate the conversation about the hypothetical person. For example, “Wang Wu is mentioned in the question. Who is Wang

Wu in your understanding?” It was commented by students that Zhang San and Wang Wu were two common names that were used to describe hypothetical characters in the mathematical examinations at the end of semesters. Wang and Zhang were understood as specific people like the subject himself/herself or a classmate. For instance, one student reported, “Wang is the same as me, except eyesight.” Wang and Zhang were also referred to as persons in a video game or in a novel. Even though the hypothetical persons were understood in different ways by students, hypothetical thinking was not a challenge for students in Grade 7, let alone the students with more education.

Table 2.5. Vignette comprehension

Potential problems	Obs	Wang: Poor vision at 5 meters ¹ (Frequency)	Zhang: Good vision at 20 meters ² (Frequency)
1. Difficulty in understanding hypothetical person	36	0%	0%
2. Irrelevant info on vignette character's sex	18	44%	44%
3. Technical term undefined (distance vision)	18	17%	17%
4. Challenge in vignette equivalence (i.e., distance)			
5 meters	28	50%	NA
10 meters	10	NA	60%
20 meters	16	NA	75%
5. Challenge in vignette equivalence (i.e., ambiguous phrases)			
Appear blurry	18	33%	NA
Angry	18	NA	22%
Familiar people's face	18	NA	11%
6. Lengthy wording	18	0%	11%

Note: By design, vignette Wang represented poor vision while vignette Zhang represented good vision.

¹ [Wang Wu] finds faces to appear blurry at a distance of 5 meters.

² [Zhang San] can recognize familiar people's faces and pick out facial expression (e.g., angry, smile) at a distance of 20 (or 10) meters quite distinctly.

C. Comprehension II: Vignette equivalence

The main challenges in vignette equivalence were the understanding of distance and the descriptions (i.e., appear blurry, familiar people's face, angry) in the vignettes. Vignette equivalence refers to the requirement that vignettes are understood in approximately the same way by all students, irrespective of their age, sex, income, education, country or other factors (Salomon, Tandon et al. 2004). These issues are discussed in depth below.

Sense of distance

The ability to accurately assess distance might vary substantially among students. In the initial vignettes, the concept of distance was introduced as '5 meters' among 28 students and '20 meters' among 18 students; in the revised vignette of Zhang, it was introduced as '10 meters' among 10 students (Table 2.5). In interviews, the sense of distance was tested by different questions. For example, "What is the length of this interview room?"; "What is your best guess of the distance between you and the eye examination chart?"; "How far is 5 meters?" The majority of students experienced difficulty in estimating the distance. The greater the actual distance was, the smaller the proportion of students who accurately estimated the distance was. Among 28 students who were asked to estimate the distance of 5 meters, only half of the students (50%) accurately estimated it and the estimation task became more challenging as the distance increased to 10 meters or 20 meters, with 40% (4 out of 10 students) and 25% (4 out of 18 students, after excluding two item non-responses) giving accurate estimations, respectively. Some of the students mentioned that it was easy to estimate the distance of 1 meter or 2 meters. The objective description of distance, e.g., 20 meters, in the vignette, introduced large variation in

understanding of the vignette character's vision. In other words, because of the difference in estimated distance, vignettes described with objective distance can hardly be interpreted equivalently among students.

It was found that the probe, "Do you know how far 5 meters are?," yielded answers that did not match with distance estimations. One student answered, "No, I don't. It is very hard to estimate the distance of 5 meters," but he estimated correctly twice regarding that distance of 5 meters (i.e., the length of the room and the distance set in the objective measure of vision). Students lacked confidence about what they knew in this case. Another student said, "I don't know how far 5 meters are." But he 'guessed' correctly or he had a good implicit sense of distance. He also had an accurate conceptualization of 10 meters. Then he recalled that, when he was in elementary school, one of the tasks he conducted was measuring length and width of the classroom.

In the second round of interviews, students were invited to compare the original questionnaire with the revised version. Students thought the version without objective distance was easier to answer and closer to students' daily life.

Vignette phrases

Subjects also reported different understandings of the described vignettes (i.e., [Wang Wu] finds faces to appear blurry at a distance of 5 meters; [Zhang San] can recognize familiar people's

faces and pick out facial expression (e.g., angry, smile) at a distance of 20 meters quite distinctly). In the first round of cognitive interviews, students reported difficulty with phrases in the vignette. For example, 33% (6 out of 18 students) exhibited difficulty with the phrase, ‘appear blurry;’ 22% (4 out of 18 students) exhibited difficulty with the phrase, ‘angry;’ and 11% (2 out of 18 students) had difficulty with the phrase, ‘familiar people's face.’

Wang’s vision was described as, “[Wang Wu] finds faces to appear blurry at a distance of 5 meters,” and ‘appear blurry’ was understood in different ways: 1) Wang cannot recognize the person; 2) Wang can recognize the person but Wang experienced double image of the face; 3) Wang cannot pick up the details such as moles on the face. Furthermore, some students wanted to know more about the details regarding the blurriness.

Subjects commented on ‘familiar people’s face’ in a richly descriptive manner in the cognitive interviewing. Eleven percent (2 out of 18 students) reported lengthy wording for the vignette involving Zhang. One student commented that it was an easier task to recognize a familiar person’s face rather than a stranger’s, and accordingly, he was struggling to rate Zhang’s vision as ‘good’ or ‘excellent.’ The term ‘familiar people’ induced more thoughts about visual input for another student. He commented that visual input was the whole person, including face and body characteristics. Further, because of the existing impression about familiar people’s other signals, such as hairstyle and walking gesture, this person can be recognized by those signals other than the face. Meanwhile, those signals from the person would also facilitate the recognition of face.

He thought the question on Zhang's vision was harder to understand than that on Wang's vision because of the length of the vignette.

Students thought observing anger on someone's face did not fit daily experience well. For instance, one student asked, "Is it possible to observe anger from facial expressions? I think anger is observed through overall body language and voice. For example, if the teacher raises his voice and waves his arm fast and firmly, I guess he is angry."

Table 2.5 summarizes all the findings regarding comprehension of vignettes. For instance, 17% of students (3 out of 18 students) had difficulty with the phrase, 'distance vision.' Besides the six problems listed in Table 2.5, among 36 students, one subject mentioned that "the same as Zhang's" can be refined as "about the same as Zhang's." In the first round of interviews, among 18 students, one subject commented that there might be no student that would rate his/her vision as "better than Zhang's" with the description of "pick out facial expression at a distance of 20 meters quite distinctly."

D. Judgment I: Reference point

It was suggested by students that a reference point was essential in the judgment of vignettes. Overall, difficulty in making judgments about the vignettes involving distance was greater in non-comparative judgment than comparative judgment. One reason for this was that subjects were using self as the reference point to make relative judgments for the

comparative condition. However, for non-comparative judgments, subjects needed to generate an absolute sense about distance or seek a new reference point for further judgment.

Interestingly, students utilized a previous vignette to make a judgment regarding the sequential vignette. For instance, the assessment of 20 meters was conducted with reference to an understanding of 5 meters. One student said, “The vignette with 5 meters helps me to understand the vignette with 20 meters. If the vignette with 20 meters is presented first, I am not sure how to answer it; maybe Zhang's vision is excellent or good. Even though I don't have the absolute sense of 20 meters, I know how far it should be, compared with 5 meters.”

Further, relevance to one's own life experience made the judgment process more engaging. Because most of the students had a good sense of their own vision, it was easier to make a judgment by comparing themselves and the vignette's protagonist. There was no standardized probe design to test whether comparative judgment or non-comparative judgment is more relevant to the students. More than 10 students mentioned that it was engaging in the comparative judgment. Meanwhile, one student found it hard to empathize with Wang Wu, and recommended rephrasing the statement to make it more engaging. He offered an example to make the vignette more relevant by suggesting, “Your classmate Wang Wu finds your face to appear blurry at a distance of 5 meters” (with suggested changes in italics).

Given that relevance to the students' daily life was an important factor to consider in designing vignettes, one challenge in comparative judgment was the limits of experience. For students with good vision, they had no or limited experiences like that of Wang's; for students with poor vision, they had a difficult time recalling the sense of good vision. In evaluating Zhang's vision in comparative judgment, one student commented that, "For the vignette with distance of 10 meters, I can hardly recall my own experience because I became near-sighted since Grade 4 and I have no memory of visual clearness at 10 meters."

E. Judgment II: Leading frame that implies a "right answer"

A leading frame might imply a "correct" answer for the participants. Eleven percent (2 out of 18 students) reported that the word 'quite' in the vignette of Zhang implied that a socially acceptable response would be 'excellent.' After reading the vignette, one student repeated, 'quite distinctively... 20 meters...' He suggested that the term 'quite distinctively' implied very good vision. He also asked, "Is it possible that a student could see very clearly at 20 meters?" He rated Zhang's vision as excellent. He had the concern that no student could have better vision than Zhang and Zhang's vision was unrealistic for him.

F. Response I: Challenges to response consistency

In order to describe the vignette scenario clearly and to ensure response consistency (Grol-Prokopczyk 2014), it was stated in the survey, "When answering the next questions, I want you to think about Wang Wu and Zhang San of your age and gender. Please think about Wang Wu

and Zhang San's vision without glasses or contact lenses. ” It was hypothesized that response consistency might be enhanced by specifying information about the vignette characters' age and gender in the preface. Response consistency requires that individuals use the response categories in a similar way in self-assessment and in evaluating hypothetical scenarios (Salomon, Tandon et al. 2004). However, in the first round of interviews, 44% (8 out of 18 students) reported that the information on the gender of vignette characters in the preface was irrelevant. The information about 'gender' was removed after the first round of interviews; and, in the second round of interviews, among 18 students, no students asked whether the hypothetical person was a girl or a boy student. In the final version, the vignette character's names were revised to little Zhang (Xiao Zhang in Chinese) and little Wang (Xiao Wang in Chinese) to omit the information about a vignette character's gender.

The assumption of response consistency may be challenged in this study by the tendency within Asian culture to be critical in self-evaluation and tolerant to others' performance. Eleven percent (4 out of 36 students) had inconsistent standards in rating their own vision and the vignette character's vision. One study found that Asians are more critical in self-assessment than North Americans, measured by larger discrepancy of actual-ideal self among Asians (Heine and Lehman 1999) and such cultural norms were uncovered through cognitive interviews during self-assessment of vision. In the case that different response standards applied in self-assessment of vision and the evaluation of vignettes, indirect comparison between self-assessment and vignettes by researchers introduces bias in assessing vision. However, comparative judgment

does not require response consistency because it is a joint evaluation between self and the vignette in a single question.

The following is an example in which a student rated his vision as “good” in self-assessment.

Student: *“In responding to survey, I don’t like to choose the extreme answer.”*

Interviewer: *“Why?”*

Student: *“It is proper to be modest. I think most people would not go for the extreme answers, such as ‘excellent’.”*

Interviewer: *“Is there any student who has excellent vision in your class?”*

Student: *“Hmm. Actually, my vision might be among the best in my class. But if I think about a greater scope, my vision is not excellent.”*

Interviewer: *“Would you elaborate a little bit more? What kind of scope are you thinking about?”*

Student: *“For example, in the entire school or in the entire Jingning County, there are so many people with excellent vision. My vision is just ‘good’.”*

Interviewer: *“How sure are you that your vision is good?”*

Student: *“In most of the cases, I would say it is good. I think my vision is good among the classmates. If you ask me to compare with another classmate, I might be the better one. But, I like to be modest.”*

The student rated Wang’s vision as ‘poor.’ He thought the question was easy to answer.

Interviewer: *“Is there a difference between the categories, ‘fair’, ‘poor’ and ‘very poor’?”*

Student: *“Not so much difference between ‘poor’ and ‘very poor’. Wang’s vision might be very poor...or poor. I am sure that I will not rate Wang’s vision is fair.”*

Interviewer: *“Really? Why did you rate his vision as ‘poor’?”*

Student: *“Probably...I should not rate others’ vision as ‘very poor’. I don’t like that extreme answer, either. I mentioned that I would not rate my own vision as ‘excellent’.”*

Interviewer: *“You seem to be very lenient to Wang but you are strict with yourself. How many students would apply the same principle?”*

Student: *“I think most of the students would. Maybe, 70% to 80%.”*

After receiving the highest possible score on the Snellen chart, the student had the chance to re-rate his vision, which he rated as “good” rather than “excellent.” He insisted on the original answer and he was very sure to choose “good” as the answer. When he was asked about the reason, he said, *“If others’ [vision] is very poor, I’d like to be lenient. I seldom rated others’ performance as ‘very poor’. For myself, I would like to be humble.”*

G. Response II: Comparison of vignettes

Within-subject comparison of two vignettes (Appendix 2.2) was conducted in the interviews and the quantitative results are presented in Table 2.6.

Table 2.6. Difficulty in responding to vignettes, by objective vision

Problems	Obs	Wang: Poor vision at 5 meters ¹ (Frequency)	Zhang: Good vision at 20 meters ² (Frequency)
Which one is more difficult to answer	22	55%	45%
With objectively measured vision 4.0-4.3	8	100%	0%
With objectively measured vision 4.4-4.7	6	67%	33%
With objectively measured vision 4.8-5.3	8	0%	100%

Note: By design, vignette Wang represented poor vision while vignette Zhang represented good vision.

¹ [Wang Wu] finds faces to appear blurry at a distance of 5 meters.

² [Zhang San] can recognize familiar people's faces and pick out facial expression (e.g., angry, smile) at a distance of 20 (or 10) meters quite distinctly.

Comparing the two vignettes, among 31 students, 29% (9 students) thought vignettes Wang and Zhang were equivalent in response difficulty. The remaining 22 students thought one was more difficult than the other. Specifically, 55% of the 22 students thought that it was more difficult to answer the vignette of Wang than that of Zhang.

After categorizing students into three groups by level of visual acuity, a clear pattern emerged. For students with poor objective vision with values from the Simplified Snellen Chart ranging from 4.0 to 4.3, all of them (8 students) thought that the vignette describing Wang (who had poorer vision) was more difficult to answer. Meanwhile, for the students with measured vision ranging from 4.8 to 5.3, all of them (8 students) reported that the vignette describing Wang was relatively easier to be answered. For the students in the middle-range of objective vision, 67% (4

out of 6 students) experienced more difficulty in answering the vignette of Wang. One student's vision was closer to Zhang's and he found it was harder to answer the question on Zhang's vision. He doubted whether he could see clearly at the distance of 20 meters. Another student, with poor objective vision, commented that the question on Zhang's vision was easier than the question on Wang's vision. She was 100% sure that her vision was worse than Zhang's vision, despite the fact that she had no sense of how far 20 meters was.

Overall, these results suggest that it was more difficult for students to evaluate the vignette that was closer to their own situation and this phenomenon occurred for both comparative judgment and non-comparative judgment.

H. Response III: Comparison of Vignette methods

Among 35 students, 17% (6 students) thought comparative judgment and non-comparative judgment were equivalent. For instance, one student commented that for Zhang's vision, the difference between comparative judgment and non-comparative judgment was not noticed until the researcher mentioned it. The remaining 29 students thought one was more difficult than the other.

Among 29 students, 79% of the students reported that non-comparative judgment was more difficult to answer and the pattern persisted among different visual acuity groups (Table 2.7).

Table 2.7. Difficulty in responding to question formats, by objective vision

		Comparative judgment (Frequency)	Non- comparative judgment (Frequency)
1. Which one is more difficult to answer	29	21%	79%
With objectively measured vision 4.0-4.3	11	27%	73%
With objectively measured vision 4.4-4.7	8	13%	88%
With objectively measured vision 4.8-5.3	8	25%	75%
2. Uncertainty of the answer	32	13%	48%

In responding to vignette questions, the uncertainty about the distance hindered the capacity of respondents to make a conclusion. In the case that the students had no sense about 20 meters or 10 meters or 5 meters, they reported uncertainty in response to vignette questions after commenting on the challenge of distance estimation.

Furthermore, difficulty in estimating distance was even more substantial in non-comparative judgment than comparative judgment, as reported by students. For students with the sense that they could only see clearly at a distance of roughly 1 meter, even though they did not know how far ‘a distance of 5 meters’ was, the lack of sense in distance did not interfere with comparative judgment. For instance, one student had no sense of how far 5 meters or 20 meters was and her logic in judgment is quoted here: “I cannot see clearly at a distance of 1 meter and it is inferred that Wang’s vision is better than mine, even though I don’t know how far 5 meters is.”

Evaluation of the vignette in five categories became challenging because of the uncertainty about how far 5 meters was. For vignette Zhang, she found she was very certain about the answers in

comparative judgment but was not certain in non-comparative judgment. One student elaborated that, because he had a good sense about his own vision, it was easier to compare the vignette character's vision with his.

Among students who reported non-comparative judgment to be easier than comparative judgment, one student's comment was representative: "It is very straightforward to rate Wang's vision. When I am asked about the comparison between Wang's vision and mine, I first think about Wang's vision, and then I evaluate my own vision. I have to think back and forth when doing the comparison, but very quickly came to conclusion when rating Wang's vision."

Comparative judgment and non-comparative judgment might yield conflicting results. One student rated her vision the same as Zhang's. When asked to rate Zhang's vision in non-comparative judgment, she rated Zhang's vision as 'excellent' while rating her own as 'good.' She had no specific sense about her own visual performance at 20 meters; however, it was her best guess that her vision was about the same as Zhang's and, after a second thought, she re-rated her vision as 'worse than Zhang's.' Further study could quantify the difference between comparative judgment and non-comparative judgment suggested by the cognitive interviewing in this study.

Among 32 students, for comparative judgment and non-comparative judgment, the uncertainty of one's answer was 13% (4 students) and 48% (15 students), respectively (Table 2.7). One reason

presumed by the researchers was that three categories were listed in comparative judgment while five categories were listed in non-comparative judgment.

I. Designing vignettes

Designing concise and yet complete vignette descriptions is clearly challenging (Kapteyn, Smith et al. 2011a) and it was suggested by students in this study that the vignette needed to be described in such a way that visual capacity was defined by what can be seen clearly and what cannot be seen clearly. “I thought more about Wang’s vision and became unsure about my answer. Can Wang see clearly at a distance of 4 meters?” “I am not sure about Wang’s vision if the only thing I know is that he is my age and has difficulty seeing clearly at a distance of 5 meters. Does he see clearly at the distance of 2 meters? How about 3 meters?” The student found it helpful to add, “But Wang can see clearly at the distance of 2 meters.”

The tradeoff of designing multiple vignettes emerged in cognitive interviews. Researchers have the tendency to design multiple vignettes to collect data about subject’s evaluation criteria; however, from students’ perspective, it was probable that some of the vignettes were redundant. For students with visual clearness less than 5 meters, after they answered the question about Wang’s vision at a distance of 5 meters, they found the question about Zhang’s vision at a distance of 10 meters either “too obvious” or irrelevant. “I can only see clearly at a distance of 2 meters. My vision is worse than Wang’s and definitely worse than Zhang’s. After answering the

question about Wang's vision, I found the sequential question about Zhang's vision not interesting at all."

Lack of attention or attention lapses might result in inconsistent ranking of multiple vignettes. For two vignettes in this study, Wang's vision was worse than Zhang's. However, one student rated Wang's vision as 'good' and Zhang's vision as 'poor.' When he was asked to elaborate the answers, he switched the answers. He confessed that he lost concentration in responding to the vignette questions and he rated the vignettes without too much thought. "I'm still distracted by my poor performance in the earlier exam, and find it hard to concentrate on answering the survey questions."

The finalized vignettes were:

In the cafeteria, [Xiao Wang] can clearly recognize students sitting at his table, but not those sitting at the next table.

From the last row in the classroom, [Xiao Zhang] can clearly recognize his teacher, but not the small written text on the blackboard.

Those vignettes were planned to be used in a sequential large-scale survey experiment (Su 2015a). Results from cognitive testing were used to generate hypotheses for further quantitative

study to quantitatively further the inquiry with statistical power sufficient to yield conclusive findings.

The following two quotes shed light on the complexity of the cognitive process in responding to vignettes.

“It is difficult to rate Wang’s vision. It depends. If we are talking about the criteria to select basketball team members, Wang’s vision is good. But in terms of performance in classroom setting, his vision is just fair. ”

“What is the surrounding Wang is in? If it is bright area, Wang's vision is poor. If it is very dark, his vision is fair.”

V. Discussion and limitations

This study documented vignette comprehension, judgment, and responses among students with different educational levels and with different objective levels of visual ability in rural China.

In the finalized vignette descriptions, only Chinese common surnames such as “Zhang” and “Wang” were used to describe the hypothetical character. These names, which are gender-neutral in a Chinese context, might provide an inherent advantage, because it eliminates the possibility of gender bias, which has been seen in other cultures. One study in the U.S. and the Netherlands

suggested that participants responded differently to vignettes with a female name than with a male name (Kapteyn, Smith et al. 2007; Jürges and Winter 2013). Therefore, vignette equivalence may not hold, at least if the potentially subtle connotations of vignette persons' names are not fully controlled (Jürges and Winter 2013). Further, it is found by Grol-Prokopczyk and colleagues that the gender of the respondent, as opposed to the gender of the vignette character, drives observed gender differences in rating style (Grol-Prokopczyk 2014) but omission of information about a vignette character's gender is not feasible in most linguistic settings (Grol-Prokopczyk, Freese et al. 2011).

Findings from this study add to a broader literature on cultural norms and psychology in an Asian context. Many of the challenges in cross-cultural survey work recognized by researchers (Johnson 1998; Johnson 2006; Willis and Miller 2011) can be at least partially attributed to the impact of culture on self-reports because culture exerts a fundamental influence on basic psychological processes (Keesing 1974; Lehman, Chiu et al. 2004; Chiu and Hong 2013). Findings from this study suggest that the cultural view of being “strict with oneself and lenient towards others” can play a large role in how students respond to vignettes and conduct self-assessment in vision. There was probably an impact of culture on the responses to vignettes, especially in non-comparative judgment. Specifically, it was found in this study that participants of cognitive interviews were hesitant to criticize others and tended to under-report their own capacity. It is very likely that this cultural norm, categorized as an “East Asian paradigm” in cross-cultural psychology studies, exists in China as well as other Asian countries (Lehman, Chiu et al. 2004). For instance, it has been documented that the Japanese are more self-critical

than North Americans (Heine and Lehman 1999) and this culture has consequences for work-related values (Hofstede 1984). Compared to European North Americans, East Asians have different patterns in thinking, perception, and self-concept (Nisbett, Peng et al. 2001). Although East Asians are capable of analytical and person-focused reasoning, they are more likely to apply the culturally encouraged way of thinking (Lehman, Chiu et al. 2004). East Asian thought systems have been found to be holistic rather than analytic (Nisbett, Peng et al. 2001). Similar to the results found by other researchers (Masuda and Nisbett 2001), in this study, participants thought holistically about the context in which the vignette characters were embedded.

The study has important limitations. First, students whose parents were migrant workers were not included in this study because of the feasibility to obtain parental consent for those under 18 years old. Thus, the findings from the interviews of 7th graders were restricted to students who have parents present. A subsequent quantitative study found that only 29% of the minor students had parents present to give consent for their dependent (Su 2015a). For adult students, the children of migrant workers might be included but no information regarding parental working status was collected to identify such students. The group of students without parents present might be marginalized or worse-off, and the cognitive process in this group of students remained unknown. Second, the sample size of the cognitive interviews was 36. With a small sample size, the quantitative results were suggestive rather than conclusive. Third, the quantitative results might not be generalized to the wider population because a representative sample was not included. Accordingly, the qualitative output, such as the direct quotes, might be more informative with the merit of in-depth interviews. Fourth, it was observed that some Chinese

students were modest in self-assessment of vision in interviews, but it is left unknown whether the students would honestly report or over-report visual acuity in a self-administered anonymous survey, in which the interviewer is absent. Fifth, due to small sample size, it was not explored regarding the systematic cognitive differences by any particular respondent characteristics, except the objective level of visual ability. Last but not least, the findings are about the domain of vision and it might be not replicated to other domains.

VI. Conclusion and implications

Overall, we conclude that anchoring vignette techniques can be used to obtain meaningful self-reports of visual function among young respondents. This was uncovered through cognitive interviewing, which revealed comprehension, judgment, and responses to the vignettes. Also, important suggestions emerged for designing vignette questions in an Asian context.

In the domain of distance vision, vignette descriptions, including the hypothetical person, were understandable for the students at the educational levels of grade seven and beyond in the domain of vision. It was suggested that participants' evaluation of vignettes was affected by knowledge of vignette characters' age but not gender. Objective distance, used in the World Health Survey, might cause difficulties in comprehension.

There are several findings regarding the practice of using anchoring vignette techniques in an Asian context. Because omitting information about a vignette character's gender is an option for linguistic reasons in Chinese, there is probably no need to include male or female pronouns or

first names with implied gender when using anchoring vignette techniques in Chinese. Response inconsistency occurred in absolute judgment and it can be attributed to values in Asian culture about humility in self-assessment and tolerance toward others.

General suggestions in vignette design emerged. It was suggested that the vignette is described in such a way that the domain of interest is defined by what can be done and what cannot be done by the vignette character. Researchers have the tendency to design multiple vignettes to collect more data; however, from participants' perspective, it is probable that some of the vignettes are redundant. The participants used self or the previous vignette as a reference point in a judgment. It was more difficult to reach an answer or to be certain about the answer in non-comparative judgment than comparative judgment.

This study revealed the value of cognitive interviewing in identifying areas for improvement in vignette design within an Asian context. Future studies are encouraged to explore vignette design in other domains and in other cultural contexts.

Competing interests

The author declares that she has no competing interests.

Human subject research ethics

This study was reviewed and approved by IRB at Harvard T. H. Chan School of Public Health (ID: IRB13-3213).

Assessing the Validity of Anchoring Vignettes in Measuring Self-rated Health:

A Survey Study in China Using Objective Vision as a Gold Standard

Abstract

Background The anchoring vignette technique has been developed to improve interpersonal comparability in self-reports and is widely used in empirical studies, including the World Health Survey. However, violation of assumptions underlying the technique might introduce biases in study results. The objective of this study was to assess the validity of three vignette methods (i.e., indirect comparison of self and vignettes, direct comparison of self and vignettes, and primed self-assessment by vignettes), using the Simplified Snellen Chart as an objective measure of true visual acuity. The null hypothesis was that there would be no significant difference in validity between the three vignette methods, with the ex-ante expectation that vignette methods were more valid than pure self-assessment.

Methods The survey experiment was conducted in high schools in rural China. Students were randomized into two groups, traditional techniques and alternative techniques, and given a survey. After completing the survey, they conducted a Snellen test to objectively measure their vision. In traditional techniques, self-assessment was conducted before the evaluation of vignette scenarios, through which the validities of pure self-assessment and indirect comparison were evaluated. In alternative techniques, the students first compared their own vision with the

vignette characters' vision and then conducted self-assessment such that self-assessment would be primed by exposure to the vignette question, through which the validities of direct comparison and primed self-assessment were evaluated. The area under a receiver operating characteristic (ROC) curve was measured to summarize sensitivity as well as specificity of vignette methods.

Results Data were collected from 4,006 students. The values of the areas under the ROC curves were 0.91, 0.90, 0.88 and 0.85 for primed self-assessment, self-assessment, direct comparison, and indirect comparison, respectively. In pairwise comparisons of self-assessment and vignette methods, the area under the ROC curve for self-assessment was significantly greater than for indirect comparison and for direct comparison. The area under the ROC curve for primed self-assessment was greater than that for self-assessment but the difference was not significant. Pairwise comparisons of vignette methods yielded statistically significant differences in validity.

Conclusion Surprisingly, vignette methods were not improvements over self-assessment in measuring distance vision. The indirect comparison technique, most commonly used in surveys to date, was the least valid technique. Although the priming effect was non-significant, primed self-assessment was the most valid technique, mainly due to the high validity of self-assessment. The results suggest that self-assessment can provide a close proxy to actual visual health.

Key words: self-assessment; anchoring vignettes; direct comparison; indirect comparison; priming effect; receiver operating characteristic (ROC); the area under the ROC curve; visual acuity; need for glasses

I. Introduction

Self-reports of health behaviors have been widely used in national health surveys, such as the National Health Interview Survey (NHIS) in the U.S. and the National Healthcare Service Survey (NHSS) in China. In addition, field experiments also heavily depend on self-reports (Brook, Ware et al. 1984; Nichols 1991; Baker and van der Gaag 1993; Newhouse 1996; Doorslaer and Jones 2003; Lindeboom and Van Doorslaer 2004; Bago d'Uva, Doorslaer et al. 2007). However, self-reports are subject to challenges of interpersonal incomparability (King, Murray et al. 2004). To address interpersonal incomparability and improve the validity of self-reports, methods have been developed such as the “unfolding” latent variable model (Javaras and Ripley 2007), the Bayesian hierarchical approach (Rossi, Gilula et al. 2001), and the anchoring vignette technique (King, Murray et al. 2004; Salomon, Tandon et al. 2004). Anchoring vignettes have been popular (Chevalier and Fielding 2011) and have been suggested as promising tools for improving interpersonal comparability of self-rated health (Grol-Prokopczyk, Freese et al. 2011).

Interpersonal incomparability was defined as two individuals who are equal on the underlying quantity of interest but nonetheless have unequal probabilities of providing the same answer in a survey inquiring about said quantity of interest (Hopkins and King 2010). For example, in one study trying to measure the prevalence of psychosis symptoms, participants were asked, “Have you had an experience of seeing visions or hearing voices that others could not see or hear when you were not half asleep, dreaming or under the influence of alcohol or drugs?” Surprisingly, the survey found that in Nepal, nearly 32% of participants responded affirmatively to this question, whereas no other country measured above 14% and the vast majority were below 5% (Nuevo,

Chatterji et al. 2010). It was found that the Nepalese interpret the act of “hearing voices that others could not hear” as a very normal, positive part of life, in which people connect with the voices of their ancestors for wisdom and guidance. In contrast, “hearing voices” was perceived as a symptom of psychotic disorder (i.e., hallucinations) by Western psychiatrists (Myers 2011).

Another example is that residents in the U.S. and the Netherlands use different response scales in self-assessment of work disability. Particularly, as some studies showed, for the same level of actual work disability, Americans were less likely to categorize themselves as disabled from work when compared to Dutch residents (Kapteyn, Smith et al. 2007; Kapteyn, Smith et al. 2011b) and other Europeans (Kapteyn, Smith et al. 2011b).

In an effort to improve interpersonal comparability, anchoring vignette techniques have been applied in numerous cross-cultural surveys, including the Survey of Health, Ageing and Retirement in Europe (SHARE), the Study on Global AGEing and Adult Health (SAGE), and the 70-country World Health Survey (WHS) (Grol-Prokopczyk 2014). Anchoring vignettes have also been widely and innovatively used to examine diverse topics such as self-rated health and political efficacy (Salomon, Tandon et al. 2004; Damacena, Vasconcellos et al. 2005; Kapteyn, Smith et al. 2007; Hopkins and King 2010; Tampubolon 2010; Van Soest, Andreyeva et al. 2011; Van Soest, Delaney et al. 2011; Kapteyn, Smith et al. 2011a; Kapteyn, Smith et al. 2011b). For example, anchoring vignettes have been applied in a novel way to address cross-country incomparability of work disability (Cutler 2011; Van Soest, Andreyeva et al. 2011).

The original form of this technique, called indirect comparison, asks survey respondents to first self-assess the domain of interest and then make an assessment for a hypothetical character presented in a vignette. It is hypothesized that by using responses to one or more vignettes, researchers can recalibrate the responses to self-assessment questions, ultimately ensuring that more valid estimations can be made. In recent years, tailored statistical models have been developed to parametrically (Wand, King et al. 2007) and non-parametrically (Wand 2013) recalibrate self-assessment.

Researchers have proposed alternative forms of anchoring vignette techniques, such as primed self-assessment by vignettes as well as direct comparison between self and vignettes. Hopkins and King found that by “priming” participants first with a vignette, and then conducting a self-assessment, participants might be able to develop a common sense of the question’s meaning before making a self-assessment (Hopkins and King 2010). These authors also tested direct comparison, which combines vignettes and self-assessments into a single direct comparison. The researchers ultimately found that this method could lead to inconsistent and less informative responses. While this research provides important insight into the potential merits of different vignette techniques, more research is needed to assess the validity of these different anchoring vignette techniques. In this study, the validity of each technique, including the indirect comparison, primed self-assessment by vignette, and direct comparison, as well as pure self-assessment, is assessed against an objective measure of visual acuity used as a gold standard in the health domain of vision.

Among different health domains, vision was chosen to validate these vignette methods mainly for three main reasons. First, there is a well-established objective measure of visual acuity, the Simplified Snellen Chart, which can act as a gold standard by which the validity of the vignette techniques can be assessed. Second, in the domain of distance vision, both objective measures and self-assessment have been used in cross-country surveys, such as the World Health Organization surveys (King, Murray et al. 2004). Third, there are financial and administrative advantages of using surveys, if surveys are proven to be a valid measure. In conducting objective measures of vision, health professionals, devices, and standardized procedures are required and objective measures can only be conducted one-by-one. However, surveys are relatively inexpensive, can be self-administered, and can be conducted simultaneously, making them a potentially cost-effective alternative to objective measures if proven valid.

In the second section of this paper, we describe the function, assumptions, and potential improvement of the most commonly used vignette method: indirect comparison. In the third section, we outline the evaluation methods used in the study. In brief, we first used a randomized survey experiment, followed by a free assessment of vision via Snellen chart to provide a gold standard measure. In the fourth section, the estimation strategy is presented. The primary metric by which the validity of each technique was assessed is the receiver operating characteristic (ROC) curve, in terms of predictive ability for the outcomes of a measured visual acuity test. In the fifth section, the results are presented. Specifically, we find that, from least to

most valid, the techniques are indirect comparison, direct comparison, self-assessment, and primed self-assessment. In the last section, concluding remarks are presented.

II. Indirect comparison in anchoring vignettes

In the health domain of distance vision, interpersonal incomparability refers to the situation when students with the same visual capacity have different self-assessed vision. To address this interpersonal incomparability, an anchoring vignette technique can be used. The following is an example of an anchoring vignette, with a focus on design logic.

A. Function

When traditional techniques are used, as in the case of the World Health Survey, survey participants are first given a standard self-assessment question, and then they are introduced to hypothetical individuals in multiple vignettes. Given the focus on distance vision in this study, we will use it as the basis for an example.

For this example, imagine two hypothetical respondents, John and Anne. One vignette tells respondents: From the last row in the classroom, [Zhang] can clearly recognize his teacher, but not the small written text on the blackboard. The survey then asks: “Would you say [Zhang]’s distance eyesight is ‘excellent’, ‘good’, ‘fair’, ‘poor’, ‘very poor’?”

As seen in Figure 3.1, if John indicates that he himself has “good vision,” and that Zhang has “excellent vision,” we know that John has poorer vision than Zhang does through indirect comparison of self-assessment and the response to the vignette. If Anne reports having “fair vision” but also says that Zhang has “fair vision,” we know that her vision is similar to Zhang’s vision and thus better than that of John’s vision. Zhang’s level of vision becomes a fixed anchor on the scale, allowing researchers to correct for interpersonal incomparability by relating each respondent’s vision to Zhang.

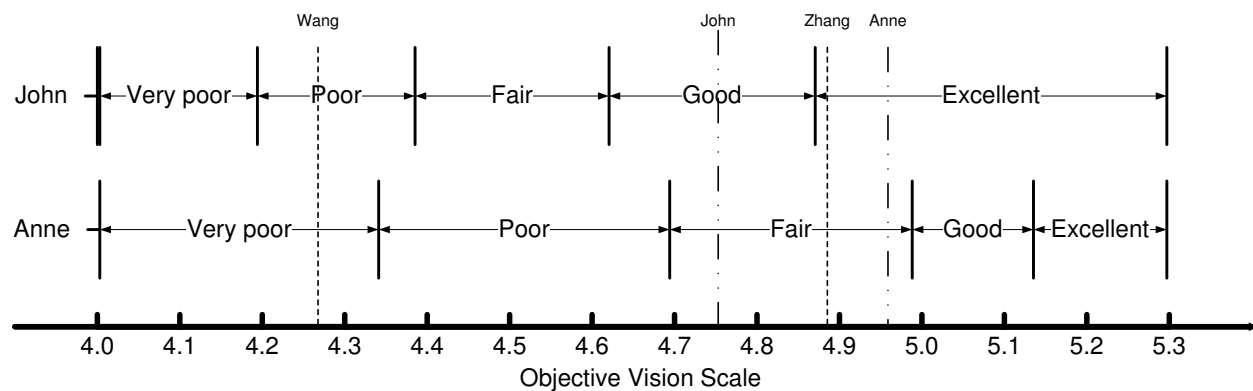


Figure 3.1. Response scales and vignettes to calibrate self-reports

However, if only the self-assessment were considered, it would seem that John’s vision was better than Anne’s. Because there can be significant variation in self-perception of vision at the individual level, respondents may assess their own vision using different scales.

B. Assumptions of indirect comparison

There are two important assumptions of indirect comparison: vignette equivalence and response consistency (Salomon, Tandon et al. 2004). Vignette equivalence refers to the requirement that the underlying domain of interest represented in each vignette is understood in approximately the same way by all participants, irrespective of their age, sex, income, education, or other factors (Salomon, Tandon et al. 2004). Response consistency refers to the requirement that participants use the response categories in a similar way when evaluating hypothetical scenarios as in self-assessments (Salomon, Tandon et al. 2004).

Testing of the assumptions and performance of anchoring vignettes has yielded mixed results. For instance, in one study examining self-assessed health, the assumption of vignette equivalence was upheld (Rice, Robone et al. 2011). However, in another study, it was documented that objective distance (e.g., 5 meters) and ambiguous vignette phrases might cause difficulty in comprehension in the domain of distance vision (Su 2015b). This situation can lead to a violation in vignette equivalence. In that same study, some respondents applied different standards in self-assessment and in the assessment of others due to an Asian cultural tendency of being “strict with oneself and lenient towards others” (Su 2015b). This cultural tendency could lead to a violation in response consistency. For instance, one study found that both vignette equivalence and response consistency were violated in Asian countries (i.e., Bangladesh, Indonesia, Vietnam) in the domains of mobility and cognition (Hirve, Gomez-Olive et al. 2013). In other studies, in-depth interviews were conducted with participants following survey completion to understand response consistency (Kapteyn, Smith et al. 2011a; Au and Lorgelly 2014) and it was found that

response consistency was satisfied for EQ-5D-5L (Au and Lorgelly 2014) and the domain of sleep (Kapteyn, Smith et al. 2011a), but it was not satisfied in other health domains, such as affect, pain, and cardiovascular diseases (Kapteyn, Smith et al. 2011a).

Violation of either assumption might introduce bias, making the technique less valid. It is important to keep in mind these underlying assumptions, as they are important in the discussion about why certain methods may have been shown to be more valid than others.

C. Improving anchoring vignettes

While the indirect comparison technique has been the predominant vignette method used in survey research, efforts have been made toward developing and evaluating different forms of anchoring vignettes. Two alternative vignette methods, direct comparison and primed self-assessment by vignettes, have been explored (Hopkins and King 2010).

Direct comparison is a joint evaluation, in which respondents are asked to evaluate their own vision compared to a hypothetical character's vision in a vignette. Thus, this technique is free from the assumption of response consistency because only one standard can be applied in this single question (Table 3.1). However, as shown by Hopkins and King, direct comparison may induce more inconsistent answers than indirect comparison. They hypothesized that the inconsistent answers derived from the tendency of participants to choose the response most similar to the scenario in the vignette (Hopkins and King 2010).

Table 3.1. Assumptions of vignette techniques

Vignette method	Vignette equivalence	Response consistency
Indirect comparison	Yes	Yes
Direct comparison	Yes	No
Primed self-assessment	Yes	No

Another alternative is the primed self-assessment by vignettes (Hopkins and King 2010), which is also free from the assumption of response consistency (Table 3.1). While the traditional anchoring vignette technique introduces respondents to a vignette following self-assessment, the primed self-assessment by vignette reverses this ordering by starting with the vignette and then conducting a self-assessment. Because of this ordering, the vignette “primes” the self-assessment. The priming effect has been found to be significant in surveys examining a variety of topics. For instance, if a participant is first asked a question about marriage and then asked to self-assess their happiness, one study has shown that the self-assessment of happiness was impacted (i.e., primed) by the question on marriage status (McClendon and O'Brien 1988).

Because of its significant impact on self-assessment, this priming effect, and by extension the primed self-assessment method, has been condoned as a strategy to improve self-report surveys (Hopkins and King 2010). It has been stated that this priming effect should not be viewed as a bias to avoid but rather an effective means of communicating the question’s meaning (Hopkins and King 2010). In addition to these strengths, the approach of primed self-assessment does not

require the use of vignette data or the involvement of multiple vignettes. More specifically, when a vignette presents a specific scenario of a general domain (e.g., vision), it serves as a consistent reference point for all respondents in making a judgment in the self-assessment of vision. Thus, direct comparison is quite similar to primed self-assessment by vignette in the way that the student is encouraged to compare self with the vignette, using the vignette as a reference point in judgment. In this way, a vignette functions to correct interpersonal incomparability in a psychological and cognitive process. For researchers, there is no extra step to take in data analysis. In order to make a vignette function by arranging question order, the only assumption is vignette equivalence.

Hopkins and King evaluated the alternative vignette methods to improve indirect comparison in the context of political efficacy and rest/energy, for which no gold standard exists. Therefore, they relied on construct validity and discriminant validity (Hopkins and King 2010). Availability of a gold standard allows us to assess criterion validity, in terms of predictive capacity for the outcomes of a measured visual acuity test, in contrast to construct and discriminant validity (Hopkins and King 2010). In this study, three vignette techniques are validated in the domain of vision, using the Simplified Snellen chart to measure true visual acuity.

III. Experimental methods

This survey experiment was conducted in four schools in Zhejiang province in China. All students were randomized at the individual level in the survey experiment with equal probability.

Students were asked to respond to questions about visual acuity without glasses for both themselves and the hypothetical characters in vignettes. After that, each student had his or her vision objectively measured via Simplified Snellen Chart.

A. Research question and hypotheses

The research question was whether self-report surveys with anchoring vignettes could act as a proxy for objective vision. The null hypothesis was that there is no significant difference in validity between the three vignette methods, with the ex-ante expectation that vignette methods were more valid than self-assessment. The alternative hypothesis was that direct comparison or primed self-assessment was more valid than indirect comparison.

B. Procedures

Cognitive interviewing and pretesting

Before the experiment, cognitive interviewing (Tanur 1992; Schuman and Presser 1996; Sudman, Bradburn et al. 1996; Collins 2003; Willis 2005) and pretesting of vignette questionnaires were conducted to generate a sense of whether and how vignettes were understood, judged, and responded to by students in a typical, rural school. In total, 36 in-depth interviews were conducted to document how the participants would respond to different vignettes and different vignette methods (Su 2015b). Furthermore, the questionnaires were pretested among 126 students from a school that would not be included in the survey experiment. Adjustments were made according to feedback from the cognitive interviews and pretesting.

Recruitment

This study was conducted in Jingning County, Zhejiang Province, China, in September and October, 2014. The pilot was conducted among 129 students to finalize the procedures of the survey experiment and objective measure of vision. Parental consent, school presidents' consent, adult students' consent, and minor students' assent were acquired. The majority of subjects in the study population were boarding students, whose parents lived in villages or outside the county as migrant workers. Accordingly, the Institutional Review Board at Zhejiang University approved the school president to consent the survey as the supervisor of the boarding students, taking into account the minimal risk of the study and the goal of the study to serve the worse-off subgroup of students.

Totally, 5,129 students in four middle schools in Jingning County, China, were defined as the study population and 4,362 students were invited to participate in the study according to the results in power calculation. Finally, 4,320 students were recruited, with a participation rate at 99%. Among the recruited students, 4,147 students responded to the survey, with a response rate at 95%. Due to an implementation mistake, 141 students had objective measures of vision before responding to the survey and they were excluded in analysis. Finally, 4,006 students were included in data analysis.

Survey experiment

Students were randomized into two groups at the individual level: direct comparison and indirect comparison, as seen in Table 3.2. For the group exposed to traditional techniques, self-assessment was conducted before the evaluation of vignette scenarios; however, in alternative techniques, the students were first asked to compare their own vision with the vignette characters', which was followed by self-assessment. Each participant self-administered the survey.

Table 3.2. Randomization strategy

Randomized group	Survey questions
Traditional techniques	Pure self-assessment + Evaluation of vignettes
Alternative techniques	Direct comparison + primed self-assessment

Objective vision

After completing the survey, all participants received an objective measurement of their vision, conducted by three members of the Center for Disease Control (CDC) and Prevention, using the Simplified Snellen Chart. The fee for these objective measurements of vision was covered by rural community insurance and students bore no out-of-pocket payment to receive the service. Distance vision without glasses was measured even if the student had visual correction. To mitigate measurement error, objective measures of vision were conducted with the following procedures: 1) Three CDC officials were trained and the same implementers conducted objective

measures of vision in four schools; 2) The same equipment was used for all sites; 3) All students were asked to sit on chairs with fixed height for objective measures of vision to avoid the variation from leg heights, given that the variation in torsos was small; and 4) All students were encouraged to keep their response time to the simplified Snellen chart to several seconds in length. All of these procedures ensured minimal measurement error.

C. Measures

Measures in the survey experiment

The main measures in the survey experiment were self-assessment and responses to vignettes (Table 3.3). Information was also collected about clusters (school, grade, and class), whether the students wear glasses, and two demographic variables (age and sex). The vignettes were designed in ascending level of visual acuity. The distance between two tables in the cafeteria was around 2 to 3 meters in four participating schools. The distance between the last row and the front of the classroom was around 8 meters, and this was standardized at the study site.

Table 3.3. Randomized survey questionnaires

Traditional techniques	Alternative techniques
<p>1. Do you wear glasses? (1) Yes, (2) No.</p> <p><i>If you answered 'Yes', now, I would like you to take off your glasses or think about your own vision <u>without</u> glasses.</i></p> <p>2. At the present time, would you say your eyesight is: (1) Excellent, (2) Good, (3) Fair, (4) Poor, (5) Very poor.</p> <p>When answering the next questions, I want you to think about Wang and Zhang and imagine them as being your age. Please think <i>about Wang's and Zhang's visual acuity</i> without glasses.</p> <p>3. In the cafeteria, [Xiao Wang] can clearly recognize students sitting at his table, but not those sitting at the next table. Would you say [Wang]'s distance eyesight is: (1) Excellent, (2) Good, (3) Fair, (4) Poor, (5) Very poor.</p> <p>4. From the last row in the classroom, [Xiao Zhang] can clearly recognize his teacher, but not the small written text on the blackboard. Would you say [Zhang]'s distance eyesight is: (1) Excellent, (2) Good, (3) Fair, (4) Poor, (5) Very poor.</p>	<p>1. Do you wear glasses? (1) Yes, (2) No.</p> <p><i>If you answered 'Yes', now, I would like you to take off your glasses or think about your own vision <u>without</u> glasses. When answering the next questions, I want you to think about Wang and Zhang and imagine them as being age. Please think about <i>Wang's and Zhang's visual acuity without</i> glasses.</i></p> <p>2. In the cafeteria, [Xiao Wang] can clearly recognize students sitting at his table, but not those sitting at the next table. Would you say your eyesight is: (1) Better than Wang's (2) About the same as Wang's (3) Worse than Wang's</p> <p>3. From the last row in the classroom, [Xiao Zhang] can clearly recognize his teacher, but not the small written text on the blackboard. Would you say your eyesight is: (1) Better than Zhang's (2) About the same as Zhang's (3) Worse than Zhang's</p> <p>4. At the present time, would you say your eyesight is: (1) Excellent, (2) Good, (3) Fair, (4) Poor, (5) Very poor.</p>

Measure of objective vision

Objective measures of vision were gathered via Simplified Snellen Chart (please refer to Figure 3.4 in Appendix 3.2). The Simplified Snellen Chart only consists of the letter “E” with different directions, whereas the normal Snellen chart consists of more letters, such as “E”, “F”, and “T.” The Simplified Snellen Chart is used to estimate visual acuity with the respondent standing at 5 meters from the chart. The measurement ranges from 4.0 to 5.3, an arithmetic sequence with 0.1 progressions (Figure 3.4). The larger the number is, the better the respondent's vision is. In Figure 3.4, for the top three vision measures, correct responses to all letters in each vision measure were required to affirm the objective vision between 4.0 and 4.2; between the vision measure of 4.3 and the vision measure of 5.3, it was required to correctly respond to at least three “E”s in each vision measure.

The Simplified Snellen Chart is a well-established objective measure of vision that is a legitimate gold standard applicable to the study participants. The Simplified Snellen Chart was approved by the National Bureau of Standards in China to be a national standard in testing visual acuity (GB115331989). It was approved by the Ministry of Health to be used nationwide in China on March 27th, 1989. The Simplified Snellen Chart also rules out the difficulties of Snellen letters and common misidentifications (Mathew, Shah et al. 2011) and the capacity of participants to respond to this eye chart is most likely to be independent from educational level. In Figure 3.4, except for the top three vision measures, the probability of overestimating true vision by 0.1 due to chance alone is smaller than 6.25% (i.e., $1/64$). More specifically, the probability to guess the direction of one letter “E” correctly by chance is $1/4$ and the probability to guess the directions of

three “E”s correctly by chance is only 1/64. Therefore, the Simplified Snellen Chart is very likely to be a valid tool for measuring visual acuity.

IV. Estimation strategy

A. Outcomes

There were five main outcomes for this study: pure self-assessment, primed self-assessment, indirect comparison, direct comparison, and need for glasses by objective measurement.

Both pure self-assessment and primed self-assessment were gathered directly from survey responses, ranging from 1 to 5. Self-assessed vision was noted as y_i for subject i .

Equation 3.1.

$$\text{Pure or primed self – assessment} = \begin{cases} 1, & \text{if } y_i \text{ is excellent} \\ 2, & \text{if } y_i \text{ is good} \\ 3, & \text{if } y_i \text{ is fair} \\ 4, & \text{if } y_i \text{ is poor} \\ 5, & \text{if } y_i \text{ is very poor} \end{cases}$$

For both indirect comparison and direct comparison, an ordinal categorical variable is denoted as C_i for subject i . For indirect comparison, the value of C_i was synthesized from self-assessment and vignette responses (King, Murray et al. 2004; King and Wand 2007; Wand, King et al. 2007). For direct comparison, the value of C_i was directly gathered from the responses to vignette questions.

C_i has three categories, in responding to a single vignette, Zhang.

$$\text{Equation 3.2.} \quad C_i = \begin{cases} 1, & \text{if } y_i \text{ is better than Zhang's} \\ 2, & \text{if } y_i \text{ is the same as Zhang's} \\ 3, & \text{if } y_i \text{ is worse than Zhang's} \end{cases}$$

C_i has five categories, in responding to two vignettes, Zhang and Wang, in the simplest case.

Equation 3.3.

$$C_i = \begin{cases} 1, & \text{if } y_i \text{ is better than Zhang's} \\ 2, & \text{if } y_i \text{ is the same as Zhang's} \\ 3, & \text{if } y_i \text{ is better than Wang's but worse than Zhang's} \\ 4, & \text{if } y_i \text{ is the same as Wang's} \\ 5, & \text{if } y_i \text{ is worse than Wang's} \end{cases}$$

In general, the number of categories of C_i is $2J+1$ for J vignettes.

However, in responding to multiple vignettes, participants might rank the vignettes inconsistently. For instance, if a student rated his vision to be worse than Wang's and better than Zhang's, the ordinal categorical variable might be coded as '1' and '5', which indicates an inconsistent answer. Ranking consistency refers to the requirement that the participants can comprehend the ordinal ranking of vignettes consistently with the actual order of underlying vision levels in vignettes (King, Murray et al. 2004). When students consistently order vignettes,

the ordinal categorical variable has a single value. For two vignettes, Table 3.4 shows that ordinal categorical variables consisted of five consistent answers and four inconsistent answers.

Table 3.4. Values in two vignette settings and interpretation

Wang [1,3]	Zhang [1,3]	Two vignettes [1,5]	Interpretation
1	1	1	Better than Zhang's
1	2	2	The same as Zhang's
1	3	3	Worse than Zhang's and Better than Wang's
2	1	Inconsistent answer	NA
2	2	Inconsistent answer	NA
2	3	4	The same as Wang's
3	1	Inconsistent answer	NA
3	2	Inconsistent answer	NA
3	3	5	Worse than Wang's

Therefore, as shown in Equation 3.4, the complete summary is that C_i has five ordinal categories and one category denoted as ‘inconsistent answers,’ incorporating information from two vignettes, Zhang and Wang.

Equation 3.4.

$$C_i = \begin{cases} 1, & \text{if } y_i \text{ is better than Zhang's} \\ 2, & \text{if } y_i \text{ is the same as Zhang's} \\ 3, & \text{if } y_i \text{ is better than Wang's but worse than Zhang's} \\ 4, & \text{if } y_i \text{ is the same as Wang's} \\ 5, & \text{if } y_i \text{ is worse than Wang's} \\ \text{Inconsistent answer} \end{cases}$$

Both direct comparison and indirect comparison were constructed according to Equation 3.4. But, the inconsistent survey responses (see Table 3.4) need to be processed for further analysis.

The handling of the inconsistent responses could potentially be designed to make the survey method as sensitive as possible (identifying all cases where glasses are truly needed), as specific as possible (identifying respondents as needing glasses only if they truly do), or designed to balance the two competing objectives. On one extreme, sensitivity would be maximized by recoding all inconsistent answers as positive (i.e. indicating a need for glasses). At the other extreme, specificity would be maximized by recoding all inconsistent answers as negative (i.e., indicating no need for glasses).

In this study, coding of inconsistent answers in the study was intended to reach maximum sensitivity. The reason is that survey instruments are planned to be used as screening tools to identify students who need glasses. Students who are classified as positive cases by a survey will be followed up with full eye examinations to fit for glasses. In such cases, sensitivity of the survey instrument is more important than specificity, given that full eye examinations are affordable. Therefore, inconsistent answers were coded to represent the situation that self-assessed vision was worse than Wang's.

The inconsistent answers in Equation 3.4 were recoded with a value of 5 in Equation 3.5 to maximize sensitivity. After coding the inconsistent answers, the ordinal categorical variable C_i was determined using Equation 3.5 below.

Equation 3.5.

$$C_i = \begin{cases} 1, & \text{if } y_i \text{ is better than Zhang's} \\ 2, & \text{if } y_i \text{ is the same as Zhang's} \\ 3, & \text{if } y_i \text{ is better than Wang's but worse than Zhang's} \\ 4, & \text{if } y_i \text{ is the same as Wang's} \\ 5, & \text{if } y_i \text{ is worse than Wang's or inconsistent answer} \end{cases}$$

Last but not least, need for glasses by objective vision, as a gold standard, was represented as a binary variable. If the student's objectively measured vision was less than or equal to 4.6 on the Snellen chart, he or she was identified as needing glasses. Those with a score above 4.6 did not need glasses. This point is used as an official cutoff to determine whether or not the student should be given a full eye examination by the Center for Disease Control in the study county in China.

B. The areas under the ROC curves

The null hypothesis that all vignette techniques were equally valid survey tools was examined through Pearson Chi-square tests for the areas under receiver operating characteristic (ROC) curves for four survey methods: indirect comparison, direct comparison, pure self-assessment

and primed self-assessment. Six pairwise comparisons were conducted, using STATA Version 12.

Receiver operating characteristics (ROC) is a summary indicator of each method's classification ability. The ROC analysis was based on Signal Detection Theory, which was developed during World War II to analyze radar images. Signal detection theory was used to measure the ability of radar receiver operators to identify whether a blip on the screen represented an enemy target, an alliance ship, or noise. Since the 1970's, the ROC has been used in medical areas to determine the accuracy of diagnostic tools (Zweig and Campbell 1993; Pepe 2003).

Estimation of the difference between the areas of ROC curves involved a three-stage statistical procedure.

First, using objectively measured visual acuity as a gold standard, the pair of sensitivity and specificity was evaluated for each discrimination threshold that classifies students as "needing glasses" (Figure 3.1). Six pairs of sensitivity and specificity are presented with the true positive rate (i.e., sensitivity) on the vertical axis against the false positive rate (i.e., 1-specificity) on the horizontal axis (Figure 3.1). For each threshold, the closer the point is to the upper left, the greater the diagnostic accuracy. There were a total of six potential discrimination thresholds, because the responses from four survey methods each were in five ordinal categories. These six thresholds are listed below, with reference to category values in Equation 3.5.

Threshold 1: No students need glasses
 Threshold 2: Students with category value 5 need glasses
 Threshold 3: Students with category values 5 and 4 need glasses
 Threshold 4: Students with category values 5, 4 and 3 need glasses
 Threshold 5: Students with category values 5, 4, 3 and 2 need glasses
 Threshold 6: All students need glasses

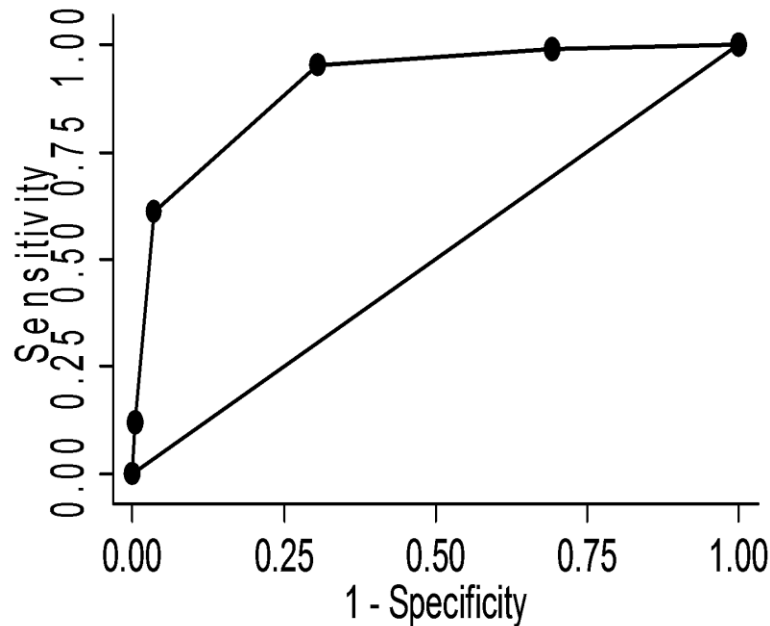


Figure 3.2. Sample of ROC curve

Second, these six points were then connected into curves, as the discrimination threshold for what values constitute a positive find is varied. Once the ROC curve is generated, the area of each curve can be measured. By construction, the value of the area under the ROC curve ranges from 0.5 to 1. When the ROC curve is a diagonal line, the value of the area under the ROC curve is 0.5. When the ROC curve along with the diagonal line forms an equilateral triangle, the value reaches 1. The areas under the ROC curves can be interpreted as the overall validity of the

survey methods as it combines information regarding both the sensitivity and specificity from all six thresholds.

Third, tests of equality of the areas under the ROC curves were conducted using pairwise chi-square tests (Metz 1978; Zweig and Campbell 1993; Cleves and Rock 2002; Pepe 2003).

V. Results

A. Descriptive characteristics

The characteristics of the study population are summarized in Table 3.5.

Table 3.5. Demographic characteristics

Variable	Obs	Mean
Age	3945	15
% male	3986	47%
% Junior high	4006	70%
% wear glasses	3996	47%
% need glasses*	4003	61%
Vision, all students*	4003	4.5
Vision, glasses wearers*	1896	4.2

* Objective vision was measured by Snellen chart. The values in Snellen chart range from 4.0 to 5.3. The larger the value in Snellen chart, the better the vision. A student was classified as needing glasses if the value of Snellen chart was less than or equal to 4.6.

B. Main results

Table 3.6 shows that there was no significant difference between the two randomized groups in terms of basic demographic characteristics. Randomization was likely to be successful.

Table 3.6. Pre-intervention demographic characteristics, by randomized group

Variable	Direct comparison	Indirect comparison	P> t
Age	15	15	0.93
% male	49%	46%	0.09
% Junior high	70%	70%	0.94
% wear glasses	47%	48%	0.83
% need glasses	61%	62%	0.73
Vision, all students	4.5	4.5	0.62
Vision, glasses wearers	4.2	4.2	0.27
Observations	2,000	2,006	

Ranking inconsistency was empirically tested and Table 3.7 shows that it occurred at 7.4% and 8.6% for direct comparison and indirect comparison, respectively. The difference between the consistency of answers in direct comparison and indirect comparison was not significant.

Table 3.7. Ranking inconsistency in direct comparison and indirect comparison

	Obs	Inconsistency	Percent
Direct comparison	1,988	147	7.4%
Indirect comparison	2,001	171	8.6%

The center-seeking tendency, the tendency for participants to select the central categorical response, was examined. As shown in Table 3.8, there is no evidence of center-seeking in direct comparison.

Table 3.8. Direct comparison: Distribution of responses

Vignette	Response	Direct comparison
Wang	Better than Wang's	65%
	The same as Wang's	25%
	Worse than Wang's	10%
Zhang	Better than Zhang's	32%
	The same as Zhang's	30%
	Worse than Zhang's	37%

ROC curves and Areas under the curves

The ROC curves for each of the four survey methods are displayed in Figure 3.3. In each plot, from left to right, the threshold ranges from 1 to 6. Each dot represents a pair of sensitivity and specificity for the threshold.

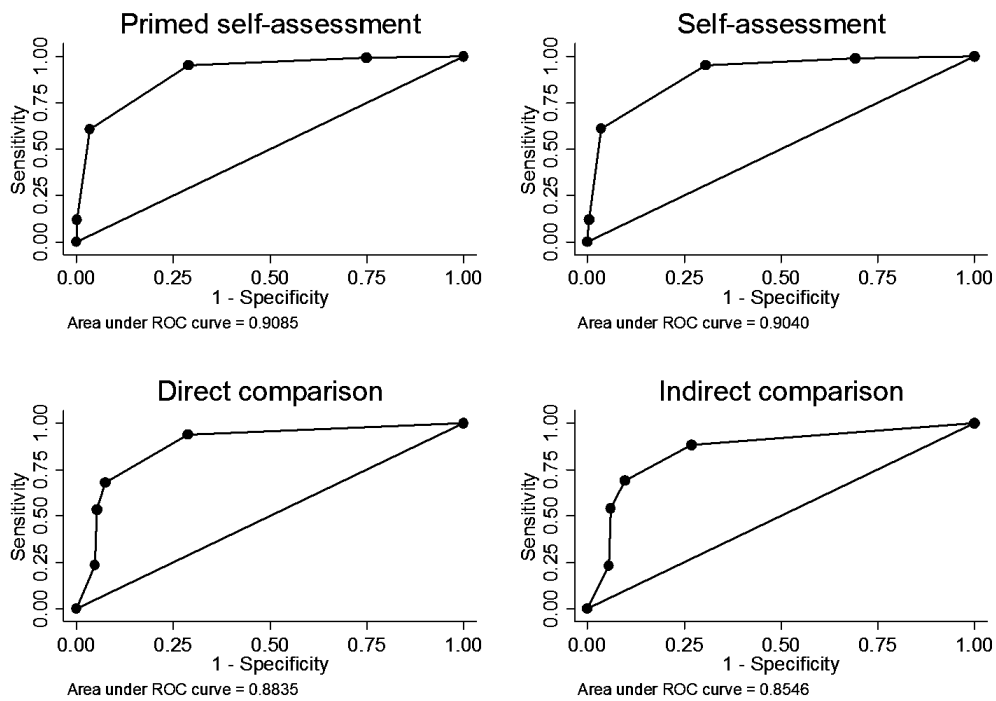


Figure 3.3. ROC curves, by survey technique

The areas under the ROC curves were 0.909, 0.904, 0.884, and 0.855, for primed self-assessment, self-assessment, direct comparison, and indirect comparison, respectively (Table 3.9).

Table 3.9. Areas under the ROC curves, by survey technique

Survey instrument	Obs	ROC			
		Area	Std. Err.	[95% Conf. Interval]	
Primed self-assessment	1991	0.909	0.006	0.90	0.92
Self-assessment	2002	0.904	0.006	0.89	0.92
Direct comparison	1985	0.884	0.008	0.87	0.90
Indirect comparison	2001	0.855	0.009	0.84	0.87

Difference of the areas under the ROC curves

In two-way comparisons of the areas under the ROC curves, the differences were 0.049, 0.021, and 0.004, for comparisons of self-assessment to indirect comparison, self-assessment to direct comparison, and self-assessment to primed self-assessment, respectively (Table 3.10). The difference between self-assessment and indirect comparison and between self-assessment and direct comparison was significant. There was no significant difference between self-assessment and primed self-assessment. Among vignette methods, the differences were 0.054, 0.029, and 0.025, for comparisons of indirect comparison to primed self-assessment, indirect comparison to direct comparison, and direct comparison to primed self-assessment, respectively (Table 3.10). The difference of pairwise comparison was statistically significant.

Table 3.10. Pearson Chi-square tests of the areas under the ROC curves

Survey methods	Comparison	Difference	Chi2	Prob>chi2
Self-assessment vs. vignette methods	Self-assessment - Indirect comparison	0.049	39.36	0.000
	Self-assessment - Direct comparison	0.021	3.91	0.048
	Primed self-assessment - Self-assessment	0.004	0.26	0.613
Among vignette methods	Primed self-assessment - Indirect comparison	0.054	24.7	0.000
	Direct comparison - Indirect comparison	0.029	5.7	0.017
	Primed self-assessment - Direct comparison	0.025	10.3	0.001

The improvement from indirect comparison to primed self-assessment was 0.054, attributed to priming effect by 0.004, which was non-significant, and attributed to using self-assessment by 0.049, which was significant (Table 3.11). The improvement from direct comparison to primed

self-assessment was 0.025, attributed to priming effect by 0.004, which was non-significant, and attributed to using self-assessment by 0.021, which was significant (Table 3.11).

Table 3.11. Improvement from priming effect and using self-assessment

Improvement	Difference	Chi2	Prob>chi2
Primed self-assessment - Indirect comparison	0.054	24.7	0.000
Primed self-assessment - Self-assessment	0.004	0.26	0.613
Self-assessment - Indirect comparison	0.049	39.36	0.000
Primed self-assessment - Direct comparison	0.025	10.3	0.001
Primed self-assessment - Self-assessment	0.004	0.26	0.613
Self-assessment - Direct comparison	0.021	3.91	0.048

VI. Discussion and limitations

Three vignette methods and the self-assessment technique were evaluated in this study by using an objective measure of vision as a gold standard. Important findings emerge regarding vignette methods and interpersonal incomparability. The results of the study show that the least to the most valid technique for assessing vision is indirect comparison, direct comparison, self-assessment, and primed self-assessment. Though many health surveys currently employ the indirect comparison technique, this technique was the least valid of all those examined in this study. The results of the study suggest that direct comparison can be used as a more valid approach than indirect comparison, partly because it is free from the assumption of response consistency. Furthermore, even though priming effect by vignettes showed positive but non-significant improvement, primed self-assessment remained the most valid vignette method, as this and other studies have shown (Hopkins and King 2010).

These results suggest, firstly, that indirect comparison is less valid than the direct comparison technique (Table 3.10). This is in contrast to findings by other researchers who found that direct comparison was not an improvement over indirect comparison (Hopkins and King 2010). Hopkins and King found significant ranking inconsistency and center-seeking tendency in direct comparison (Hopkins and King 2010). However, this study found that ranking inconsistency was non-significantly different between indirect comparison and direct comparison in the domain of distance vision (Table 3.7). There is no center-seeking tendency observed in direct comparison in this study (Table 3.8). One potential explanation for why the direct comparison technique performed better than indirect comparison in this study is that it is free from the assumption of response consistency.

Second, indirect comparison also was less valid than the technique of primed self-assessment. This result is consistent with other research conducted by Hopkins and King (Hopkins and King 2010). The statistical test for the priming effect described in Hopkins and King's paper was conducted using the measure of the percentage of participants who fell in the most efficacious category in political efficacy (Hopkins and King 2010). The priming effect tested significantly in a Chi-square test of the categorical variables (i.e., self-assessment and primed self-assessment), with a p-value 0.031, without involving the gold standard.

Third, the improvement of primed self-assessment over indirect comparison consists of two components: improvement from priming effect and improvement from using self-assessment (Table 3.11). Because the study design was crafted to conduct pairwise comparisons of primed self-assessment to indirect comparison, primed self-assessment to self-assessment, and self-assessment to indirect comparison, the mechanism of improvement in primed self-assessment could be examined. The priming effect improves validity in a positive but non-significant way. It is specified in Table 3.10 that there is no significant difference between self-assessment and primed self-assessment. The improvement of validity from using self-assessment is positive and significant (Table 3.11). Thus, the improvement in primed self-assessment is mainly attributed to using self-assessment rather than to a priming effect (Table 3.11).

Fourth, it is surprising that the results show that some vignette methods (direct and indirect comparisons) are less valid than pure self-assessment. At the outset, it was hypothesized that self-assessment would be highly subject to interpersonal incomparability. However, perhaps there is reason to believe that pure self-assessment is a valid measure, particularly in the domain of distance vision.

In addition, it is hypothesized that the two vignette methods (i.e., indirect comparison and direct comparison) are less valid than self-assessment because ranking inconsistency occurred at 8.6% and 7.4% in indirect comparison and direct comparison, respectively (Table 3.7), whereas this phenomenon is nonexistent with pure self-assessment or primed self-assessment. The

inefficiencies in a non-parametric method of indirect comparison were precisely from the information lost due to ranking inconsistency (King, Murray et al. 2004) and the same inefficiencies occurred in direct comparison as well. King et al. recommend a parametric method, called CHOPIT, to deal with these inefficiencies, assuming that most of the ranking inconsistency is due to random errors (King, Murray et al. 2004). In this study, all the inconsistent answers were re-coded to maximize the sensitivity or maximize the specificity of vignette methods. Therefore, the ineffectiveness of vignette methods is still likely to be attributed to information lost due to ranking inconsistency.

This study has several important limitations. First, vignette methods, including primed self-assessment, direct comparison, and indirect comparison, are all conditional on getting the original wording in the vignette and question right. The most basic aspects of question writing affect every other higher-level activity in survey research. Without close attention to these details, vignette equivalence may not be ensured. In this study, it is possible that an aspect of vignette wording or design was overlooked and that the requirement of vignette equivalence was not upheld. However, a serious effort was made in cognitive interviewing and pretesting to ensure that this was not the case (Su 2015b). Second, this survey experiment focused on distance vision, and therefore it may not be applicable to other health domains. Third, while the study focuses on statistically significant differences, the magnitude of differences was relatively small. Therefore, there is the question of whether the differences are substantively significant in the practice of screening students who need glasses.

Last but not least, a limitation of the study is that it relies on a single indicator of validity: the area under the ROC curve. It integrates both sensitivity and specificity, giving each equal weight. For intended purposes, it may be best to only focus on one, or to weight one higher than the other. While the area under the ROC curve measures criterion validity, there are other validities that can be measured, such as construct validity and test-retest consistency. For measuring criterion validity, we can either take the global approach, such as using the areas under the ROC curve, or use a local approach, such as using selected sensitivity as a criterion (Zweig and Campbell 1993). Comparison of the areas under the ROC curves is one way to summarize sensitivity and specificity, but there are other ways, such as comparison of selected decision thresholds to maximize effectiveness per cost. Specifically, sensitivity and specificity could be integrated in a decision analytic framework, in which the consequences of false positives and false negatives are accounted for in terms of either costs or health consequences or both, similar to the approach in cost-effectiveness analysis.

VII. Conclusion and implications

This study evaluated different survey methods by testing how accurate they predict the outcomes of a measured visual acuity test. This survey experiment contributes to the literature by validating vignette methods and self-assessment against objectively measured health, which has not been done before, to our knowledge. Several important conclusions emerge from the study.

First, the indirect comparison technique was the least valid among four survey instruments and the finding was significant. Because this technique is widely used in major surveys, including the World Health Survey, this result is notable. Validity of vignette methods can be significantly improved by using primed self-assessment or direct comparison. Second, there was a positive but non-significant priming effect from ordering the vignette question before self-assessment. This suggests that participants are capable of establishing a standard scale in the domain of distance vision and can conduct an accurate self-assessment without the presence of vignettes before self-assessment. Primed self-assessment allows for more valid results compared to the indirect method because of the high validity of self-assessment. Third, two vignette methods – direct and indirect comparison – were surprisingly not as valid as pure self-assessment.

Based on this study, the concerns that past uses of self-assessed vision were compromised by interpersonal incomparability appear to be unwarranted. It has been suggested that different measurement techniques should be applied to distinctive health dimensions (Guralnik, Simonsick et al. 1994), and the results from this study suggest that survey methods, particularly primed self-assessment and pure self-assessment, can provide a close proxy to actual visual health. If the financial and administrative costs of conducting objective measures of vision are a concern in screening students for visual correction, survey methods can serve as a relatively inexpensive, valid screening tool to identify students for further eye examinations.

Competing interests

The author declares that she has no competing interests.

Human subject research ethics

This study was reviewed and approved by RB at School of Public Health, Zhejiang University, China (ID: ZGL201408-1).

Clinical trial registration

The study proposal was documented via clinical trial registry (ID: NCT02244060).

Appendices

Appendix 1.1. Paper I: Design of survey instruments, pilot, and power calculation

Pretesting to design survey instruments (n=39), cognitive interviews (n=8), and pilot (n=54) were conducted before the survey experiment. Pretesting was conducted in Xi'an Jiaotong University on Nov. 8th, 2013. In total, 39 freshmen were enrolled for pretesting, with two focuses: potential recall issues and the design of statements in the list experiment.

Intravenous infusion – Students were asked about intravenous infusion use in different time frames. It was shown that 23% of freshmen had limited understanding about intravenous infusion. Accordingly, it was added that “intravenous infusion is commonly known as ‘dripping infusion’.” According to Table 1.7, about a quarter of freshmen could not recall the utilization of intravenous infusion in elementary school. The percentage dropped to 5% for recalling in senior high school. This was consistent with the intuition that it was more challenging to recall remote events compared to more recent events. According to the estimations in Table 1.7, recalled intravenous infusion use declined over time, from elementary school to senior high school. It was most reasonable and feasible to ask about the use of intravenous infusion in Grade 12 because it bore the least recall difficulty.

Table 1.7. Intravenous infusion use among students

	% among all participants	% cannot recall	% among the recalled
Elementary school	56	23	73
Junior high school	67	8	72
Senior high school	55	5	58
Grade 12	43	10	48

Non-key statements and placebo statement -- In pretesting, 12 statements were designed as the candidates for the list experiment. These were:

1. Did you do any household work in grade 12?
2. Did you read love novels in grade 12?
3. Did you read knight novels in grade 12?
4. Did you like team sports in grade 12?
5. Were TV programs about nature your favorite in grade 12?
6. Did you prefer pop music to traditional music in grade 12?
7. Did you like calligraphy in grade 12?
8. Could you see the blackboard clearly from the last row in the classroom in grade 12?
9. Did your family have a house in Hong Kong while you were in grade 12?
10. Was math your favorite course in grade 12?
11. Did you prefer word puzzles to numeric puzzles in grade 12?
12. Were you a communist party member in grade 12?

The mean of each statement and the correlation of statements were estimated to select non-key statements and a placebo question. The statement about party membership in XJU was not a good candidate for a list experiment as only 5% of the students were party members in Grade 12. Meanwhile, the test for the statement about ownership of a house in Hong Kong showed a mean of 0 and no variation. It was a good choice as a placebo question. Pearson correlations were conducted for the variable of interest (i.e., intravenous infusion in grade 12) and the ten remaining candidate statements. Freshmen who liked pop music were more likely to use intravenous infusions. Preference of pop music was negatively associated with preference of calligraphy. The preference of math was negatively associated with the preference of word puzzles; however, the correlation was not statistically significant.

In sum, according to the feedback from pretesting, the students would be asked about the use of intravenous infusion in Grade 12, in a large-scale randomized survey experiment. Four non-key statements were chosen for the list experiment, i.e., preference of classical music, preference of calligraphy, preference of math course and preference of word puzzle. House ownership in Hong Kong was chosen as the placebo.

A. Cognitive Interviewing

Cognitive interviews were conducted with a focus on the comprehension, judgment and response to the list-based question. The first round of cognitive interviews was conducted among four students in XJU Medical School, November 24th, 2013. Concurrent cognitive debriefing was

conducted without specific probes. One of the participants circled all of the statements available to her, including use of intravenous infusions, showing that she was comfortable revealing her choices. The participant said that she would prefer being asked to directly circle the specific statements on the list. She complained that the instructions suggested that she had to count all the statements that applied to her in order to answer the question about total number of statements. Two participants indicated that the statement, 'my family owns a house in Hong Kong,' was surprising. One participant wondered why the researcher wanted to know this piece of information. The statement was changed to the following: "I have visited Gaoxiong, a city in southern Taiwan." The overall survey was commented as, "too simple to be true." One of the participants was wondering what kind of research could be done with such a simple survey. Another participant declared that this was the simplest survey he had experienced.

The second round of cognitive interviews was conducted among four students in XJU Medical School, between December 30th, 2013 and January 2nd, 2014. Retrospective cognitive debriefing was applied, in which all students first answered the questionnaire and then they were invited to think aloud about the process of surveying in a private setting. Specific probes were designed for cognitive debriefing.

Probe 1: "How did you reach the answer in the list question?"

The list questions in Table 1.8 were tested through thinking-aloud.

Table 1.8. The list questions

Version	Question	Answer
A	<p>How many of the following statements were true for you in Grade 12? (Please indicate the total number but not which ones in particular.)</p> <ul style="list-style-type: none"> • Among all courses in Grade 12, my favorite was math. • I preferred pop music to classical music in Grade 12. • I visited Gaoxiong, a city in Southern Taiwan, in Grade 12. • I liked calligraphy in Grade 12. • I preferred word puzzles to numeric puzzles in Grade 12. 	<p>0 true statement</p> <p>1 true statements</p> <p>2 true statements</p> <p>3 true statements</p> <p>4 true statements</p> <p>5 true statements</p>
B	<p>How many of the following statements were true for you in Grade 12? (Please indicate the total number but not which ones in particular.)</p> <ul style="list-style-type: none"> • Among all courses in Grade 12, my favorite was math. • I preferred pop music to classical music in Grade 12. • I had intravenous infusion, commonly known as ‘dripping infusion’, in Grade 12. • I liked calligraphy in Grade 12. • I preferred word puzzles to numeric puzzles in Grade 12. 	<p>0 true statement</p> <p>1 true statements</p> <p>2 true statements</p> <p>3 true statements</p> <p>4 true statements</p> <p>5 true statements</p>

Student #1 and Student #2 had no problem with the question. They specifically commented on each statement and elaborated the reason why it did or did not apply to them. Student #3 recommended to change the first statement from, “Among all courses in Grade 12, my favorite was math,” to, “I prefer math course to Chinese course in Grade 12,” to reduce the cognitive difficulty to complete comparison of all courses. She commented that the revised statement matched better with the fifth statement.

Student #3 commented that she had no exposure to music in Grade 12 and the statement did not apply to her. Student #4 said that only one statement applied to him and it was a straightforward question to him. Student #4 circled two answers in the list-based question. When it was pointed

out that she selected “1 true statement” and “4 true statements,” she explained that she misunderstood it as, “1st statement is true,” and, “4th statement is true,” because the answers were parallel to the statements in the Chinese version of the survey. She confessed that she did not pay any attention to the sentence, “Please indicate the total number but not which ones in particular.” She read through all the statements and circled two answers. She recommended re-formatting the answer as “____ true statements” for participants to fill out.

Probe 2: “What is the purpose of this study?”

It was the first time for all students to experience a list experiment. Students had no idea about the purpose of the study. When the complementary survey questionnaire was shown and the purpose of the design was introduced, those four students commented that it was an interesting way to survey intravenous infusion use. Student #2 had no idea about the purpose of the study, and, after a second thought, he said that maybe it was about the folk exchange between mainland China and Taiwan. Student #3 thought this was about whether a student was rational or more emotional in judgment and she pretended to be rational (e.g., like math course, like numerical puzzles).

Thinking-aloud: “Would you please talk a little bit more about intravenous infusion?”

Student #1 specified that he answered “yes” to the question about intravenous infusion use because he recalled that he had a severe illness in Grade 12 and used intravenous infusion.

Student #2 said that he had no specific memory of intravenous infusion in Grade 12 and his best

guess was that he probably had experience with it. Student #2 commented that intravenous infusion was not sensitive for him and he would like to reveal the true answer even if he was asked about this topic directly. When the topic was substituted with sexual behavior, he said that he would prefer to skip the question. Student #3 claimed that she had no intravenous infusion in Grade 12. For Student #4, his mother was a physician and he had sufficient knowledge about intravenous infusion.

The first three main adjustments made to the list-based question were based on the feedback from cognitive interviews and the last two adjustments were based on feedback from the research committee at Harvard. Therefore, all the statements were more coherent and closer to student life.

1. The instruction, “Please indicate the total number but not which ones in particular,” was bolded.
2. The non-key statement, “Among all courses in Grade 12, my favorite was math,” was changed into, “I prefer math course to Chinese course in Grade 12”.
3. The answer format was changed from circling a number to writing down the number. The range of the answer was specified and the examples of answer 0 and 5 were given.
4. The non-key statements were changed from preference to actual behaviors.
5. The statement, ‘I fell asleep during class at least once in Grade 12,’ was added to make the statements on intravenous infusions and smoking less obvious.

B. Pilot and Power Calculation

Using the finalized version of the questionnaire, the pilot was launched and 54 students were recruited in March, 2014. The estimated prevalence of intravenous infusions was 29% and 39% from direct questioning and indirect questioning, respectively. The estimated prevalence of smoking was 7% and 43% from direct questioning and indirect questioning, respectively. The DID was 26%. According to the power calculation, the number of enrolled participants needed was around 1,250 adult students (Table 1.9), with 80% power, alpha of 0.05, and using a one-sided test. Assuming that 5% of the students in universities were under 18 years old and the participation rate was 95%, it was planned to screen around 1,385 students in order to enroll 1,250 adult students.

Table 1.9. Results from power calculations

Sample size	Powers: Control list in DirectQ _{smoking}	Powers: Control list in DirectQ _{IV}
500	43%	40%
600	49%	46%
1000	71%	67%
1250	80%	77%
1500	87%	84%
2000	94%	92%
2500	98%	97%

Appendix 1.2. Paper I: Construction of dependent variables

The mathematical procedures involved in constructing dependent variable Y_{ji} are presented in this appendix.

The regression specification is:

$$Y_{ji} = \beta_{j0} + \beta_{j1} IV_{DirectQ} + \beta_{j2} Smoking_{List} + \beta_{j3} IV_{List} + \varepsilon_{ji},$$

in which j indicates the different constructions of dependent variable.

The data for regression are summarized in Table 1.1. The list-based responses were discrete data, ranging from 0 to 5. The direct responses were binary data, and particularly, the direct responses to the placebo question were with mean close to zero by design. The mathematical equations are presented in Table 1.2.

A. Indirect estimates of prevalence: Smoking and intravenous infusion use

$$\begin{aligned} & \text{Prevalence}_{\text{IndirectSmoking}} \\ &= \text{mean} (List_{\text{Smoking}} - List_{\text{ControlPooled}} + Direct_{\text{PlaceboPooled}}) \\ &= \text{mean} (List_{\text{Smoking}} - \frac{List_{\text{Control1}} + List_{\text{Control2}}}{2} + \frac{Direct_{\text{Placebo1}} + Direct_{\text{Placebo2}}}{2}) \\ &= - \text{mean} (\frac{List_{\text{Control1}}}{2}) - \text{mean} (\frac{List_{\text{Control2}}}{2}) + \text{mean} (\frac{Direct_{\text{Placebo1}}}{2}) + \text{mean} (List_{\text{Smoking}} + \frac{Direct_{\text{Placebo2}}}{2}) \end{aligned}$$

Equation 1.1.

Y_{1i} is constructed to estimate prevalence of smoking from the list experiment in the following manner:

$$Y_{1i} = \begin{cases} \frac{List_i}{2}, & \text{if Smoking}_{DirectQ} = 1 \\ \frac{List_i}{2}, & \text{if IV}_{DirectQ} = 1 \\ List_i + \frac{Direct_i}{2}, & \text{if Smoking}_{List} = 1 \\ \frac{Direct_i}{2}, & \text{if IV}_{List} = 1 \end{cases}$$

In which, $List_i$ is the response from a list-based question and $Direct_i$ is the response from a direct question presented in Table 1.1.

Therefore, Equation 1.1, with re-arrangement, becomes

$$\begin{aligned} & \text{Prevalence}_{IndirectSmoking} \\ &= - (Y_{1i} | \text{Smoking}_{DirectQ}=1) - (Y_{1i} | \text{IV}_{DirectQ}=1) + (Y_{1i} | \text{Smoking}_{List}=1) + (Y_{1i} | \text{IV}_{List}=1) \\ &= - \beta_{10} - (\beta_{10} + \beta_{11}) + (\beta_{10} + \beta_{12}) + (\beta_{10} + \beta_{13}) \\ &= - \beta_{11} + \beta_{12} + \beta_{13} \end{aligned}$$

Similarly, Y_{2i} is constructed to estimate prevalence of intravenous infusion use from the list experiment in the following manner:

$$Y_{2i} = \begin{cases} \frac{List_i}{2}, & \text{if Smoking}_{DirectQ} = 1 \\ \frac{List_i}{2}, & \text{if IV}_{DirectQ} = 1 \\ \frac{Direct_i}{2}, & \text{if Smoking}_{List} = 1 \\ List_i + \frac{Direct_i}{2}, & \text{if IV}_{List} = 1 \end{cases}$$

$$\text{Prevalence}_{IndirectIV} = - \beta_{21} + \beta_{22} + \beta_{23}$$

B. Difference of prevalence levels: Smoking and intravenous infusion use

$$\begin{aligned} & \text{Difference}_{Smoking} \\ &= \text{mean} (List_{Smoking} - List_{ControlPooled} + Direct_{PlaceboPooled}) - \text{mean} (Direct_{Smoking}) \\ &= \text{mean} (List_{Smoking} - \frac{List_{Control1} + List_{Control2}}{2} + \frac{Direct_{Placebo1} + Direct_{Placebo2}}{2}) - \text{mean} (Direct_{Smoking}) \end{aligned}$$

$$= - \text{mean} \left(\frac{\text{ListControl1}}{2} + \text{DirectSmoking} \right) - \text{mean} \left(\frac{\text{ListControl2}}{2} \right) + \text{mean} \left(\text{ListSmoking} + \frac{\text{DirectPlacebo1}}{2} \right) + \text{mean} \left(\frac{\text{DirectPlacebo2}}{2} \right)$$

Equation 1.2.

Y_{3i} is constructed to estimate the difference of prevalence levels from indirect questioning and direct questioning, for smoking, in the following manner:

$$Y_{3i} = \begin{cases} \frac{\text{List}_i}{2} + \text{Direct}_i, & \text{if } \text{Smoking}_{\text{DirectQ}} = 1 \\ \frac{\text{List}_i}{2}, & \text{if } \text{IV}_{\text{DirectQ}} = 1 \\ \text{List}_i + \frac{\text{Direct}_i}{2}, & \text{if } \text{Smoking}_{\text{List}} = 1 \\ \frac{\text{Direct}_i}{2}, & \text{if } \text{IV}_{\text{List}} = 1 \end{cases}$$

Therefore, Equation 1.2, with re-arrangement, becomes

Difference_{Smoking}

$$\begin{aligned} &= - (Y_{3i} \mid \text{Smoking}_{\text{DirectQ}}=1) - (Y_{3i} \mid \text{IV}_{\text{DirectQ}}=1) + (Y_{3i} \mid \text{Smoking}_{\text{List}}=1) + (Y_{3i} \mid \text{IV}_{\text{List}}=1) \\ &= - \beta_{30} - (\beta_{30} + \beta_{31}) + (\beta_{30} + \beta_{32}) + (\beta_{30} + \beta_{33}) \\ &= - \beta_{31} + \beta_{32} + \beta_{33} \end{aligned}$$

Similarly, Y_{4i} is constructed to estimate the difference of prevalence levels from indirect questioning and direct questioning, for intravenous infusion use, in the following manner:

$$Y_{4i} = \begin{cases} \frac{\text{List}_i}{2}, & \text{if } \text{Smoking}_{\text{DirectQ}} = 1 \\ \frac{\text{List}_i}{2} + \text{Direct}_i, & \text{if } \text{IV}_{\text{DirectQ}} = 1 \\ \frac{\text{Direct}_i}{2}, & \text{if } \text{Smoking}_{\text{List}} = 1 \\ \text{List}_i + \frac{\text{Direct}_i}{2}, & \text{if } \text{IV}_{\text{List}} = 1 \end{cases}$$

$$\text{Difference}_{\text{IV}} = - \beta_{41} + \beta_{42} + \beta_{43}$$

C. Difference-in-differences

DID

$$\begin{aligned}
 &= \text{Difference}_{\text{Smoking}} - \text{Difference}_{\text{IV}} \\
 &= [\text{mean}(\text{List}_{\text{Smoking}}) - \text{mean}(\text{Direct}_{\text{Smoking}})] - [\text{mean}(\text{List}_{\text{IV}}) - \text{mean}(\text{Direct}_{\text{IV}})] \\
 &= -\text{mean}(\text{Direct}_{\text{Smoking}}) + \text{mean}(\text{Direct}_{\text{IV}}) + \text{mean}(\text{List}_{\text{Smoking}}) - \text{mean}(\text{List}_{\text{IV}}) \\
 &\text{Equation 1.3.}
 \end{aligned}$$

Y_{5i} is constructed to estimating the difference-in-differences in the following manner:

$$Y_{5i} = \begin{cases} \text{List}_i, & \text{if } \text{Smoking}_{\text{List}} = 1 \text{ or } \text{IV}_{\text{List}} = 1 \\ \text{Direct}_i, & \text{if } \text{Smoking}_{\text{DirectQ}} = 1 \text{ or } \text{IV}_{\text{DirectQ}} = 1 \end{cases}$$

Therefore, Equation 1.3, with re-arrangement, becomes

DID

$$\begin{aligned}
 &= - (Y_{5i} \mid \text{Smoking}_{\text{DirectQ}}=1) + (Y_{5i} \mid \text{IV}_{\text{DirectQ}}=1) + (Y_{5i} \mid \text{Smoking}_{\text{List}}=1) - (Y_{5i} \mid \text{IV}_{\text{List}}=1) \\
 &= -\beta_{50} + (\beta_{50} + \beta_{51}) + (\beta_{50} + \beta_{52}) - (\beta_{50} + \beta_{53}) \\
 &= \beta_{51} + \beta_{52} - \beta_{53}
 \end{aligned}$$

Y_{1i}, Y_{2i}, Y_{3i} , and Y_{4i} were all best fit by Gamma distributions. Therefore, the variance function used a Gamma model and the link function was Log in MLE. In estimating the difference-in-differences, Y_{5i} was under a Negative Binomial distribution. Therefore, the variance function used a Negative Binomial model and the link function was Log in MLE. For all five estimated outcomes, the p-values are generated by the “lincom” command in STATA version 12.

Appendix 2.1. Paper II: Question appraisal to identify potential source of error in survey

INSTRUCTIONS. Use one column for EACH question to be reviewed. In reviewing each question:

1) TYPE IN QUESTION INITIALS IN THE TABLE HEAD OF THE QAS FORM.

NJ1: Non-comparative judgment, the first question

NJ2: Non-comparative judgment, the second question

CJ1: Comparative judgment, the first question

CJ2: Comparative judgment, the second question

2) Proceed THROUGH THE FORM - Enter 1 if the problem exists and 0 if there is no problem for each Problem Type (1a - 7b).

3) Whenever 1 is entered, write detailed notes on it that describes the problem.

Table 2.8. Question appraisal steps

		NJ1	NJ2	CJ1	CJ2
STEP 1 - INSTRUCTIONS: Look for problems with any introductions, instructions, or explanations from the respondent's point of view.					
1a.	CONFLICT OR INACCURACY	0	0	0	0
1b.	COMPLICATENESS	0	0	0	0
STEP 2 - CLARITY: Identify problems related to communicating the intent or meaning of the question to the respondent.					
2a.	WORDING: Question is lengthy, awkward, ungrammatical, or contains complicated syntax.	0	1	0	1
2b.	TECHNICAL TERMS are undefined, unclear, or complex.	1	1	1	1
2c.	VAGUE: There are multiple ways to interpret the question or to decide what is to be included or excluded.	1	0	1	0
2d.	REFERENCE PERIODS are missing, not well specified, or in conflict.	0	0	0	0
STEP 3 - ASSUMPTIONS: Determine whether there are problems with assumptions made or the underlying logic.					
3a.	INAPPROPRIATE ASSUMPTIONS are made about the respondent or about his/her living situation.	0	1	0	1
3b.	ASSUMES CONSTANT BEHAVIOR or experience for situations that vary.	0	0	0	0
3c.	DOUBLE-BARRELED: Contains more than one implicit question.	0	0	0	0
STEP 4 - KNOWLEDGE/MEMORY: Check whether respondents are likely to not know or have trouble remembering information.					
4a.	KNOWLEDGE may not exist	1	1	1	1
4b.	COMPUTATION problem: The question requires a difficult mental calculation.	0	1	0	1
STEP 5 - SENSITIVITY/BIAS: Assess questions for sensitive nature or wording, and for bias.					
5a.	SENSITIVE CONTENT (general): embarrassing	0	0	0	0
5b.	SENSITIVE WORDING (specific)	0	0	0	0
5c.	SOCIAL ACCEPTABLE response is implied by the question.	0	1	0	1

Table 2.8. Question appraisal steps (Continued)

QAS steps		NJ1	NJ2	CJ1	CJ2
STEP 6 - RESPONSE CATEGORIES: Assess the adequacy of the range of responses to be recorded.					
6a.	OPEN-ENDED QUESTION	0	0	0	0
6b.	MISMATCH between question and response categories.	0	0	0	0
6c.	TECHNICAL TERMS are undefined, unclear, or complex.	0	0	0	0
6d.	VAGUE response categories are subject to multiple interpretations.	1	1	0	0
6e.	OVERLAPPING response categories.	0	0	0	0
6f.	MISSING eligible responses in response categories.	0	0	0	0
6g.	ILLOGICAL ORDER of response categories.	0	0	0	0
STEP 7 - OTHER PROBLEMS not identified in Steps 1-6					
7a.	Ordering or context problems across questions.	0	0	0	0
	Sum	4	7	3	6

Appendix 2.2. Paper II: Cognitive interview protocol

Condition A (Non-comparative judgment)

Date: _____ ID# _____ Start time: _____

(Note: The same serial number is used in the consent/assent form for the same subject. For example, for Xiaomin Wu, the parental consent is with ID P1 and the assent is with ID S1, and the ID number in the interview is 1 for Xiaomin Wu.)

Instructions to be read to subject:

I'd like you to think aloud as you answer the questions. Please tell me everything you are thinking about as you go about answering them. At times I will ask you more questions about the terms or phrases in the questions and what you think a question is asking about. I'll take notes and record the interview. Don't hesitate to speak up whenever something seems unclear, is hard to answer, or doesn't seem to apply to you. We want you to be thoughtful and there is no hurry in giving an answer.

Do you have any questions before we start?

Think-aloud practice:

Let's begin with a practice question. Please think aloud as you answer.

Try to visualize the place where you live, and think about how many windows there are in that place. As you count up the windows, tell me what you are seeing and thinking about.

Now, I would like you to think about your own vision without glasses or contact lenses.

A1. At the present time, how would you describe your distance eyesight?

S1. How comfortable are you talking about your vision in the interview?

S2. What is your understanding of the phrase "vision without glasses or contact lenses"?

S3. What does "distance eyesight" mean in your own words?

There are five categories designed for this question on distance eyesight.

- (1) Excellent,
- (2) Good,
- (3) Fair,
- (4) Poor,
- (5) Very poor.

S4. Which category do you fall into?

- S5. How do you arrive at that answer? Would you walk me through your answer?
- S6. In what circumstance would you rate your vision as [excellent]? (Explore the understanding of different categories other than the one the subject falls in. This is one way to generate vignettes.)
- S7. How sure are you that [***]?
- S8. Is there a difference between the categories?
- S9. Please respond with a 0-10 scale, in which 0 represents the worst vision and 10 represents best vision. How would you rate your vision?
- 0 1 2 3 4 5 6 7 8 9 10

When answering the next questions, I want you to think about Wang Wu and Zhang San's vision without glasses or contact lenses. Wang Wu and Zhang San are both your age and gender.

- A2. [Wang Wu] finds faces to appear blurry at a distance of 5 meters. Would you say [Wang Wu]'s distance eyesight is:
- (1) Excellent,
 - (2) Good,
 - (3) Fair,
 - (4) Poor,
 - (5) Very poor.

- ICV1. How hard it is to answer this question? Why?
- ICV2. Wang Wu is mentioned in the question. Who is Wang Wu in your understanding?
- ICV3. How far is "a distance of 5 meters"?
- ICV4. How do you arrive at that answer?
- ICV5. How sure are you that[***]?
- ICV6. Is there a difference between the categories?

The previous questions asked you to talk about your vision and then to imagine and talk about the vision of another person. I would like to know how you would respond if I asked you a more direct comparison. For example, consider [Wang Wu] finds faces to appear blurry at a distance of 5 meters.

- ICV7. Would you say your distance eyesight is:
- (1) Better than Wang Wu's
 - (2) The same as Wang Wu's
 - (3) Worse than Wang Wu's
- ICV8. Did you find this question easy to answer?

A3. [Zhang San] can recognize familiar people's faces and pick out facial expression (e.g., *angry, smile*) at a distance of 20 meters quite distinctly. Would you say [Zhang San]'s distance eyesight is:

- (1) Excellent,
- (2) Good,
- (3) Fair,
- (4) Poor,
- (5) Very poor.

ICV9. How hard is it to answer this question? Why?

ICV10. Zhang San is mentioned in the question. Who is Zhang San in your understanding?

ICV11. How far is "a distance of 20 meters"?

ICV12. How do you arrive at that answer?

ICV13. How sure are you that[***]?

ICV14. Is there a difference between the categories?

ICV15. Would you say your distance eyesight is:

- (1) Better than Zhang San's
- (2) The same as Zhang San's
- (3) Worse than Zhang San's

ICV16. Did you find this question easy to answer?

Thank you for thinking aloud and speaking up on vision. We've now finished with the in-depth interview. Lastly, we would like to collect some basic information from you as well.

1. Grade: _____

2. Age: _____

3. Gender:

- (1) Male
- (2) Female

4. Do you wear glasses?

- (1) Yes
- (2) No

5. Do you wear contact lenses?

- (1) Yes
- (2) No

6. [If Yes to question 4 or question 5] How many years have you worn glasses or contact lenses? _____years

7. Did you have any eye examination in 2013?

- (1) Yes

(2) No

8. [If Yes to question 7] Do you remember what the eye examiner told you your objective vision was in your last eye examination?

(1) Yes, _____

(2) No

Measured visual acuity: _____

ICV17. [If the answers from S4 is inconsistent with objectively measured vision] You said that your vision is [***] but the eye doctor might say that you have [***] vision; why did you say that your vision is [***]?

Do you have any comments on the questionnaire?

Question for the interviewer:

9. Did the interviewee devote sufficient mental effort to answering the question thoughtfully?

(1) Yes,

(2) No.

End time: _____

Condition B (Comparative judgment)

Date: _____ ID# _____ Start time: _____

(Note: The same serial number is used in the consent/assent form for the same subject. For example, for Xiaomin Wu, the parental consent is with ID P1 and the assent is with ID S1, and the ID number in the interview is 1 for Xiaomin Wu.)

Instructions to be read to subject:

I'd like you to think aloud as you answer the questions. Please tell me everything you are thinking about as you go about answering them. At times I will ask you more questions about the terms or phrases in the questions and what you think a question is asking about. I'll take notes and record the interview. Don't hesitate to speak up whenever something seems unclear, is hard to answer, or doesn't seem to apply to you. We want you to be thoughtful and there is no hurry in giving an answer.

Do you have any questions before we start?

Think-aloud practice:

Let's begin with a practice question. Please think aloud as you answer.

Try to visualize the place where you live, and think about how many windows there are in that place. As you count up the windows, tell me what you are seeing and thinking about.

Now, I would like you to think about your own vision without glasses or contact lenses.

B1. At the present time, how would you describe your distance eyesight?

S1. Are you comfortable talking about vision in the interview?

S2. What is your understanding of "vision without glasses or contact lenses"?

S3. What does "distance eyesight" mean in your own words?

There are five categories designed for this question on distance eyesight.

- (1) Excellent,
- (2) Good,
- (3) Fair,
- (4) Poor,
- (5) Very poor.

S4. Which category do you fall into?

S5. How do you arrive at that answer? Would you walk me through your answer?

S6. In what circumstance would you rate your vision as [excellent]? (Explore the understanding of different categories other than the one the subject falls in. This is one way to generate vignettes.)

S7. How sure are you that [***]?

S8. Is there a difference between the categories?

S9. Please respond with a 0-10 scale, in which 0 represents the worst vision and 10 represents best vision. How would you rate your vision?

0 1 2 3 4 5 6 7 8 9 10

When answering the next questions, I want you to think about Wang Wu and Zhang San's vision without glasses or contact lenses. Wang Wu and Zhang San are both your age and gender.

B2. [Wang Wu] finds faces to appear blurry at a distance of 5 meters. Would you say your distance eyesight is:

- (1) Better than Wang Wu's
- (2) The same as Wang Wu's
- (3) Worse than Wang Wu's

DCV1. How hard is it to answer this question? Why?

DCV2. Wang Wu is mentioned in the question. Who is Wang Wu in your understanding?

DCV3. How far is "a distance of 5 meters"?

DCV4. How do you arrive at that answer?

DCV5. How sure are you that[***]?

DCV6. Is there a difference between the categories?

DCV7. Would you say [Wang Wu]'s distance eyesight is:

- (1) Excellent,
- (2) Good,
- (3) Fair,
- (4) Poor,
- (5) Very poor.

DCV8. Did you find this question easy to be answered?

B3. [Zhang San] can pick out details on the blackboard at a distance of 10 meters quite distinctly. Would you say your distance eyesight is:

- (1) Better than Zhang San's
- (2) The same as Zhang San's
- (3) Worse than Zhang San's

DCV9. How hard is it to answer this question? Why?

DCV10. Zhang San is mentioned in the question. Who is Zhang San in your understanding?

DCV11. How far is "a distance of 20 meters"?

DCV12. How do you arrive at that answer?

DCV13. How sure are you that[***]?

DCV14. Is there a difference between the categories?

DCV15. Would you say [Zhang San]'s distance eyesight is:

- (1) Excellent,
- (2) Good,
- (3) Fair,

- (4) Poor,
- (5) Very poor.

DCV16. Did you find this question easy to be answered?

Thank you for thinking aloud and speaking up on vision. We've now finished with the in-depth interview. Lastly, we would like to collect some basic information from you as well.

1. Grade: _____
2. Age: _____
3. Gender:
 - (1) Male
 - (2) Female
4. Do you wear glasses?
 - (1) Yes
 - (2) No
5. Do you wear contact lenses?
 - (1) Yes
 - (2) No
6. [If Yes to question 4 or question 5] How many years have you worn glasses or contact lenses? _____ years
7. Did you have any eye examination in 2013?
 - (1) Yes
 - (2) No
8. [If Yes to question 7] Do you remember what the eye examiner told you your objective vision was in your last eye examination?
 - (1) Yes, _____
 - (2) No

Measured visual acuity: _____

DCV17. [If the answers from S4 is inconsistent with objectively measured vision] You said that your vision is [***] but the eye doctor might say that you have [***] vision; why did you say that your vision is [***]?

Do you have any comments on the questionnaire?

Question for the interviewer:

9. Did the interviewee devote sufficient mental effort to answering the question thoughtfully?
 - (3) Yes,
 - (4) No.

End time: _____

Condition A (Non-comparative judgment)

1. Do you wear glasses?

- (1) Yes,
- (2) No.

If you answered 'Yes', now, I would like you to take off your glasses or think about your own vision without glasses.

2. At the present time, would you say your eyesight is:

- (1) Excellent,
- (2) Good,
- (3) Fair,
- (4) Poor,
- (5) Very poor.

When answering the next questions, I want you to think about Wang and Zhang and imagine them as being your age. Please think about Wang's *and* Zhang's *vision without glasses*.

3. In the cafeteria, [Xiao Wang] can clearly recognize students sitting at his table, but not those sitting at the next table. Would you say [Wang]'s distance eyesight is:

- (1) Excellent,
- (2) Good,
- (3) Fair,
- (4) Poor,
- (5) Very poor.

4. From the last row in the classroom, [Xiao Zhang] can clearly recognize his teacher, but not the small written texts on the blackboard. Would you say [Zhang]'s distance eyesight is:

- (1) Excellent,
- (2) Good,
- (3) Fair,
- (4) Poor,
- (5) Very poor.

Condition B (Comparative judgment)

1. Do you wear glasses?

- (1) Yes,
- (2) No.

If you answered 'Yes', now, I would like you to take off your glasses or think about your own vision without glasses.

2. At the present time, would you say your eyesight is:

- (1) Excellent,
- (2) Good,
- (3) Fair,
- (4) Poor,
- (5) Very poor.

When answering the next questions, I want you to think about Wang and Zhang and imagine them as being age. Please think about Wang's and Zhang's vision without glasses.

3. In the cafeteria, [Xiao Wang] can clearly recognize students sitting at his table, but not those sitting at the next table. Would you say your eyesight is:

- (1) Better than Wang's
- (2) About the same as Wang's
- (3) Worse than Wang's

4. From the last row in the classroom, [Xiao Zhang] can clearly recognize his teacher, but not the small written texts on the blackboard. Would you say your eyesight is:

- (1) Better than Zhang's
- (2) About the same as Zhang's
- (3) Worse than Zhang's

Appendix 3.1. Paper III: Pilot and power calculation

From the pilot, estimated areas under the ROC curves ranged from 0.87 to 0.95, which was used for power calculations. Between three groups, we made three possible two-way comparisons of the areas under the ROC curves: indirect comparison to direct comparison, indirect comparison to primed self-assessment, and direct comparison to primed self-assessment. We found that a sample size of 3,538 is required to identify an improvement of 0.03, with 80% power, alpha of 0.05, and using a one-sided test. We planned to recruit around 4,350 students, allowing us to lose 18% of students due to non-response in survey or objective measure of vision.

For survey experiment, 4,006 students were recruited. Among them, 207 students (5%) were older than 18 years and they offered consent by themselves. For the remaining 3,799 students, 1,100 students (29%) obtained parental consent and 2,699 students (71%) acquired permission from the school's president to attend the study. All students under 18 years offered assent.

Appendix 3.2. Paper III: Simplified Snellen chart

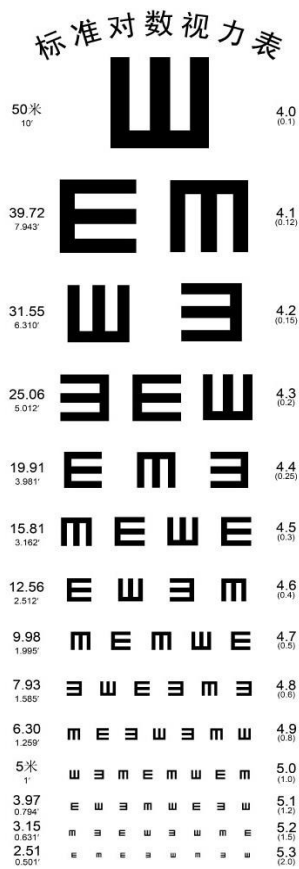


Figure 3.4. Simplified Snellen chart

Bibliography

Ahart, A. M. and P. R. Sackett (2004). "A new method of examining relationships between individual difference measures and sensitive behavior criteria: Evaluating the unmatched count technique." Organizational Research Methods 7(1): 101-114.

Antin, J. and A. Shaw (2012). Social desirability bias and self-reports of motivation: a study of amazon mechanical turk in the US and India. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM.

Au, N. and P. K. Lorgelly (2014). "Anchoring vignettes for health comparisons: an analysis of response consistency." Quality of Life Research 23(6): 1721-1731.

Bago d'Uva, T., E. K. A. Doorslaer, et al. (2007). "Does reporting heterogeneity bias the measurement of health disparities?" Health Economics: 351.

Baker, J. and J. van der Gaag (1993). "Equity in health care and health care financing: evidence from five developing countries." Equity in the finance and delivery of health care: An international perspective.

Baker, R., S. J. Blumberg, et al. (2010). "Research Synthesis AAPOR Report on Online Panels." Public Opinion Quarterly 74(4): 711-781.

Belli, R. F., M. W. Traugott, et al. (1999). "Reducing vote overreporting in surveys: Social desirability, memory failure, and source monitoring." Public Opinion Quarterly: 90-108.

Biemer, P. P., B. K. Jordan, et al. (2005). "A test of the item count methodology for estimating cocaine use prevalence." Evaluating and improving methods used in the national survey on drug use and health 149.

Blair, G. and K. Imai (2012). "Statistical analysis of list experiments." Political Analysis 20(1): 47-77.

Boeije, H. and G. Willis (2013). "The Cognitive Interviewing Reporting Framework (CIRF)." Methodology: European Journal of Research Methods for the Behavioral and Social Sciences 9(3): 87-95.

Brook, R. H., J. E. Ware, et al. (1984). "The effect of coinsurance on the health of adults." RAND Corporation, Santa Monica, Calif.

Buckley, J. (2008). "Survey context effects in anchoring vignettes." URL <http://polmeth.wustl.edu/media/Paper/surveyartifacts.pdf>.

Burden, B. C. (2000). "Voter turnout and the national election studies." Political Analysis **8**(4): 389-398.

CDC, U. (1994). "Preventing Tobacco Use Among Young People: A Report of the Surgeon General." Atlanta, Ga: US Dept of Health and Human Services.

Chevalier, A. and A. Fielding (2011). "An introduction to anchoring vignettes." Journal of the Royal Statistical Society: Series A (Statistics in Society) **174**(3): 569-574.

Chiu, C.-Y. and Y.-Y. Hong (2013). Social psychology of culture, Psychology Press.

Cleves, M. A. and L. Rock (2002). "From the help desk: Comparing areas under receiver operating characteristic curves from two or more probit or logit models." The Stata Journal **2**(3): 301-313.

Coffman, K. B., L. C. Coffman, et al. (2013). The size of the lgbt population and the magnitude of anti-gay sentiment are substantially underestimated, National Bureau of Economic Research.

Collins, D. (2003). "Pretesting survey instruments: an overview of cognitive methods." Quality of Life Research **12**(3): 229-238.

Comşa, M. and C. Postelnicu (2013). "Measuring Social Desirability Effects on Self-Reported Turnout Using the Item Count Technique." International Journal of Public Opinion Research **25**(2): 153-172.

Corstange, D. (2009). "Sensitive questions, truthful answers? Modeling the list experiment with LISTIT." Political Analysis **17**(1): 45-63.

Crowne, D. P. and D. Marlowe (1960). "A new scale of social desirability independent of psychopathology." Journal of consulting psychology **24**(4): 349.

Currie, J., W. Lin, et al. (2011). "Patient knowledge and antibiotic abuse: Evidence from an audit study in China." Journal of Health Economics **30**(5): 933-949.

Cutler, D. M. (2011). Comment on " Self-Reported Disability and Reference Groups". Investigations in the Economics of Aging, University of Chicago Press: 265-266.

d'Uva, T. B., O. O'Donnell, et al. (2008). "Differential health reporting by education level and its impact on the measurement of health inequalities among older Europeans." International journal of epidemiology **37**(6): 1375-1383.

Dalton, D. R., J. C. Wimbush, et al. (1994). "Using the unmatched count technique (UCT) to estimate base rates for sensitive behavior." Personnel Psychology **47**(4): 817-829.

Damacena, G. N., M. T. L. Vasconcellos, et al. (2005). "Perception of health state and the use of vignettes to calibrate for socioeconomic status: results of the World Health Survey in Brazil, 2003." Cadernos de Saúde Pública **21**: S65-S77.

De Jong, M. G., R. Pieters, et al. (2010). "Reducing social desirability bias through item randomized response: An application to measure underreported desires." Journal of Marketing Research **47**(1): 14-27.

Doorslaer, E. and A. M. Jones (2003). "Inequalities in self-reported health: validation of a new approach to measurement." Journal of Health Economics **22**(1): 61-87.

Droitcour, J., R. A. Caspar, et al. (1991). "The item count technique as a method of indirect questioning: A review of its development and a case study application." Measurement errors in surveys: 185-210.

Ericsson, K. A. and H. A. Simon (1980). "Verbal reports as data." Psychological review **87**(3): 215.

Falck, R., H. A. Siegal, et al. (1992). "The validity of injection drug users' self-reported use of opiates and cocaine." Journal of Drug Issues.

Farrall, S., C. Priede, et al. (2012). "Using cognitive interviews to refine translated survey questions: an example from a cross-national crime survey." International Journal of Social Research Methodology **15**(6): 467-483.

Fendrich, M. and C. M. Vaughn (1994). "Diminished lifetime substance use over time: An inquiry into differential underreporting." Public Opinion Quarterly **58**(1): 96-123.

Fisher, R. J. (1993). "Social desirability bias and the validity of indirect questioning." Journal of Consumer Research: 303-315.

Forsyth, B. H. and J. T. Lessler (1991). "Cognitive laboratory methods: A taxonomy." Measurement errors in surveys: 393-418.

Glynn, A. N. (2013). "What can we learn with statistical truth serum? design and analysis of the list experiment." Public Opinion Quarterly **77**(S1): 159-172.

Grol-Prokopczyk, H. (2014). Age and Sex Effects in Anchoring Vignette Studies: Methodological and Empirical Contributions. Survey research methods, NIH Public Access.

Grol-Prokopczyk, H., J. Freese, et al. (2011). "Using anchoring vignettes to assess group differences in general self-rated health." Journal of health and social behavior **52**(2): 246-261.

Guralnik, J. M., E. M. Simonsick, et al. (1994). "A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission." Journal of Gerontology **49**(2): M85.

Heine, S. J. and D. R. Lehman (1999). "Culture, self-discrepancies, and self-satisfaction." Personality and Social Psychology Bulletin **25**(8): 915-925.

Hesketh, T. and W. X. Zhu (1997). "Health in China. The healthcare market." BMJ **314**(7094): 1616-1618.

Hirve, S., X. Gomez-Olive, et al. (2013). "Use of anchoring vignettes to evaluate health reporting behavior amongst adults aged 50 years and above in Africa and Asia—testing assumptions." Global health action **6**.

Hofstede, G. (1984). Culture's consequences: International differences in work-related values, sage.

Holbrook, A. L., M. C. Green, et al. (2003). "Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias." Public Opinion Quarterly **67**(1): 79-125.

Holbrook, A. L. and J. A. Krosnick (2010). "Social desirability bias in voter turnout reports tests using the item count technique." Public Opinion Quarterly **74**(1): 37-67.

Hopkins, D. J. and G. King (2010). "Improving Anchoring Vignettes: Designing surveys to correct interpersonal incomparability." Public Opinion Quarterly **74**(2): 201.

Hu, S., X. Liu, et al. (2003). "Assessment of antibiotic prescription in hospitalised patients at a Chinese university hospital." Journal of infection **46**(3): 161-163.

Huang, L.-L., J. F. Thrasher, et al. (2014). "Impact of the 'Giving Cigarettes is Giving Harm' campaign on knowledge and attitudes of Chinese smokers." Tobacco control: tobaccocontrol-2013-051475.

Jamison, J. and D. Karlan (2011). Measuring Preferences and Predicting Outcomes. Symposium on Economic Experiments in Developing Countries, December.

Javaras, K. N. and B. D. Ripley (2007). "An "unfolding" latent variable model for Likert attitude data: Drawing inferences adjusted for response style." Journal of the American Statistical Association **102**(478): 454-463.

Johnson, T. P. (1998). "Approaches to equivalence in cross-cultural and cross-national survey research." ZUMA-Nachrichten spezial **3**: 1-40.

Johnson, T. P. (2006). "Methods and frameworks for crosscultural measurement." Medical Care **44**(11): S17-S20.

Johnson, T. P., J. G. Hougland, et al. (1989). "Obtaining reports of sensitive behavior: A comparison of substance use reports from telephone and face-to-face interviews." Social Science Quarterly **70**(1): 1974-1983.

Jones, E. F. and J. D. Forrest (1992). "Underreporting of abortion in surveys of US women: 1976 to 1988." Demography **29**(1): 113-126.

Jürges, H. and J. Winter (2013). "Are anchoring vignettes ratings sensitive to vignette age and sex?" Health Economics **22**(1): 1-13.

Kaminska, O. and T. Foulsham (2013). Understanding sources of social desirability bias in different modes: evidence from eye-tracking, Institute for Social and Economic Research.

Kapteyn, A., J. Smith, et al. (2011a). Anchoring vignettes and response consistency. Working Paper, RAND Center for the Study of Aging.

Kapteyn, A., J. P. Smith, et al. (2007). "Vignettes and self-reports of work disability in the United States and the Netherlands." The American Economic Review **97**(1): 461-473.

Kapteyn, A., J. P. Smith, et al. (2011b). Work disability, work, and justification bias in Europe and the United States. Explorations in the Economics of Aging, University of Chicago Press: 269-312.

Keesing, R. M. (1974). "Theories of culture." Annual review of anthropology: 73-97.

King, G., C. J. L. Murray, et al. (2004). "Enhancing the validity and cross-cultural comparability of measurement in survey research." American Political Science Review **98**(01): 191-207.

King, G. and J. Wand (2007). "Comparing incomparable survey responses: new tools for anchoring vignettes." Political Analysis **15**(1): 46-66.

King, M. F. and G. C. Bruner (2000). "Social desirability bias: A neglected aspect of validity testing." Psychology & Marketing **17**(2): 79-103.

King, N. B., S. Harper, et al. (2012). "Who cares about health inequalities? Cross-country evidence from the World Health Survey." Health Policy and Planning.

Kuklinski, J. H., M. D. Cobb, et al. (1997). "Racial attitudes and the "New South"." The Journal of Politics **59**(02): 323-349.

LaBrie, J. W. and M. Earleywine (2000). "Sexual risk behaviors and alcohol: Higher base rates revealed using the unmatched-count technique." Journal of Sex Research **37**(4): 321-326.

Lam, T., S. Chung, et al. (1997). Tobacco advertisements were associated with positive attitudes to smoking in children who had never smoked. 10th World Conference on Tobacco or Health, Beijing, China, August 24-28, 1997.

Lee, S., N. Schwarz, et al. (2014). "Culture-sensitive question order effects of self-rated health between older Hispanic and non-Hispanic adults in the United States." Journal of aging and health **26**(5): 860-883.

Lehman, D. R., C.-y. Chiu, et al. (2004). "Psychology and culture." Annu. Rev. Psychol. **55**: 689-714.

Li, Q., J. Hsia, et al. (2011). "Prevalence of smoking in China in 2010." New England Journal of Medicine **364**(25): 2469-2470.

Lindeboom, M. and E. Van Doorslaer (2004). "Cut-point shift and index shift in self-reported health." Journal of Health Economics **23**(6): 1083-1099.

Liu, C., J. Xiao, et al. (2003). "A compromise between self-enhancement and honesty: Chinese self-evaluations on social desirability scales." Psychological reports **92**(1): 291-298.

Martinez, M. D. (2003). "Comment on "Voter turnout and the national election studies"." Political Analysis **11**(2): 187-192.

Masuda, T. and R. E. Nisbett (2001). "Attending holistically versus analytically: comparing the context sensitivity of Japanese and Americans." Journal of Personality and Social Psychology **81**(5): 922.

Mathew, J. A., S. A. Shah, et al. (2011). "Varying Difficulty of Snellen Letters and Common Errors in Amblyopic and Fellow Eyes." Archives of ophthalmology **129**(2): 184.

McClendon, M. J. and D. J. O'Brien (1988). "Question-order effects on the determinants of subjective well-being." Public Opinion Quarterly **52**(3): 351-364.

McElrath, K., R. Dunham, et al. (1995). "Validity of self-reported cocaine and opiate use among arrestees in five cities." Journal of Criminal Justice **23**(6): 531-540.

McNaghy, S. E. and R. M. Parker (1992). "High prevalence of recent cocaine use and the unreliability of patient self-report in an inner-city walk-in clinic." JAMA: the journal of the American Medical Association **267**(8): 1106-1108.

Meng, T., J. Pan, et al. (2014). "Conditional Receptivity to Citizen Participation: Evidence from a Survey Experiment in China."

Metz, C. E. (1978). Basic principles of ROC analysis. Seminars in nuclear medicine, Elsevier.

Miller, K., D. Mont, et al. (2011). "Results of a cross-national structured cognitive interviewing protocol to test measures of disability." Quality & quantity **45**(4): 801-815.

MOH, C. (2006). Somking and Health. 10th World Conference on Tobacco or Health, Washington DC.

Myers, N. L. (2011). "Update: Schizophrenia across cultures." Current Psychiatry Reports **13**(4): 305-311.

Nápoles-Springer, A. M., J. Santoyo-Olsson, et al. (2006). "Using cognitive interviews to develop surveys in diverse populations." Medical Care **44**(11): S21-S30.

Nápoles-Springer, A. M. and A. L. Stewart (2006). "Overview of qualitative methods in research with diverse populations: Making research reflect the population." Medical Care: S5-S9.

Nederhof, A. J. (1985). "Methods of coping with social desirability bias: A review." European Journal of Social Psychology **15**(3): 263-280.

Newhouse, J. P., Ed. (1996). Free for All?: Lessons from the RAND health insurance experiment. Cambridge, MA, Harvard University Press.

Nichols, P. (1991). Social survey methods: a fieldguide for development workers, Oxfam.

Nisbett, R. E., K. Peng, et al. (2001). "Culture and systems of thought: holistic versus analytic cognition." Psychological review **108**(2): 291.

Nuevo, R., S. Chatterji, et al. (2010). "The continuum of psychotic symptoms in the general population: a cross-national study." Schizophrenia bulletin: sbq099.

Pasick, R. J., S. L. Stewart, et al. (2001). "Quality of data in multiethnic health surveys." Public Health Reports **116**(Suppl 1): 223.

Paulhus, D. (1988). "Balanced inventory of desirable responding (BIDR)." Acceptance and Commitment Therapy. Measures Package **41**.

Pepe, M. S. (2003). The statistical evaluation of medical tests for classification and prediction, Oxford University Press.

Ploughman, M., M. Austin, et al. (2010). "Applying cognitive debriefing to pre-test patient-reported outcomes in older people with multiple sclerosis." Quality of Life Research **19**(4): 483-487.

Presser, S. and L. Stinson (1998). "Data collection mode and social desirability bias in self-reported religious attendance." American Sociological Review: 137-145.

Raghavarao, D. and W. T. Federer (1979). "Block total response as an alternative to the randomized response method in surveys." Journal of the Royal Statistical Society. Series B (Methodological): 40-45.

Rice, N., S. Robone, et al. (2011). "Analysis of the validity of the vignette approach to correct for heterogeneity in reporting health system responsiveness." The European Journal of Health Economics **12**(2): 141-162.

Rossi, P. E., Z. Gilula, et al. (2001). "Overcoming scale usage heterogeneity: A Bayesian hierarchical approach." Journal of the American Statistical Association **96**(453): 20-31.

Salomon, J. A., A. Tandon, et al. (2004). "Comparability of self rated health: cross sectional multi-country survey using anchoring vignettes." BMJ **328**(7434): 258.

Schuman, H. and S. Presser (1996). Questions and answers in attitude surveys: Experiments on question form, wording, and context, Sage Publications, Inc.

Su, Y. (2015a). Assessing the Validity of Anchoring Vignettes in Measuring Self-rated Health: A Survey Study in China Using Objective Vision as a Gold Standard. Department of Global Health and Population. Boston, Harvard University. **Doctor of Science**.

Su, Y. (2015b). Designing Vignettes and Question Formats to Measure Distance Vision: Evidence from Cognitive Interviews among Students in China. Department of Global Health and Population. Boston, Harvard University. **Doctor of Science**.

Sudman, S., N. M. Bradburn, et al. (1996). Thinking about answers: The application of cognitive processes to survey methodology, Jossey-Bass.

Sun, X., S. Jackson, et al. (2009). "Prescribing behaviour of village doctors under China's New Cooperative Medical Scheme." Social Science & Medicine **68**(10): 1775-1779.

Tampubolon, G. (2010). "Multilevel anchoring vignettes for comparative study of health inequalities in 48 countries of the WHO."

Tanur, J. M. (1992). Questions about questions: inquiries into the cognitive bases of surveys, Russell Sage Foundation Publications.

Tourangeau, R., L. J. Rips, et al. (2000). The psychology of survey response, Cambridge University Press.

Tsuchiya, T. (2005). "Domain estimators for the item count technique." Survey Methodology **31**(1): 41-51.

Tsuchiya, T., Y. Hirai, et al. (2007). "A study of the properties of the item count technique." Public Opinion Quarterly **71**(2): 253-272.

Van Soest, A., T. Andreyeva, et al. (2011). Self-reported disability and reference groups. Investigations in the Economics of Aging, University of Chicago Press: 237-264.

Van Soest, A., L. Delaney, et al. (2011). "Validating the use of anchoring vignettes for the correction of response scale differences in subjective questions." Journal of the Royal Statistical Society: Series A (Statistics in Society) **174**(3): 575-595.

Wada, K., R. Kakuma, et al. (2011). "Factors Associated With Preferences for Health System Goals in Japan A Pilot Study of the World Health Survey." Asia-Pacific Journal of Public Health **23**(5): 721-729.

Wand, J. (2013). "Credible Comparisons Using Interpersonally Incomparable Data: Nonparametric Scales with Anchoring Vignettes." American Journal of Political Science **57**(1): 249-262.

Wand, J., G. King, et al. (2007). "Anchors: Software for Anchoring Vignette Data." Journal of Statistical Software.

WHO (2005). "Pocket book of hospital care for children: Guidelines for the management of common illnesses with limited resources." **First Edition**.

WHO (2011). "WHO report on the global tobacco epidemic, 2011: warning about the dangers of tobacco: executive summary."

WHO (2013a). "Pocket Book of Hospital Care for Children: Guidelines for the Management of Common Childhood Illnesses " **Second Edition**.

WHO (2013b). WHO report on the global tobacco epidemic, 2013: enforcing bans on tobacco advertising, promotion and sponsorship, World Health Organization.

Willis, G., D. Lawrence, et al. (2008). "Translation of a tobacco survey into Spanish and Asian languages: the Tobacco Use Supplement to the Current Population Survey." Nicotine & tobacco research **10**(6): 1075-1084.

Willis, G. B. (2005). Cognitive interviewing: A tool for improving questionnaire design, Sage Publications, Incorporated.

Willis, G. B. and K. Miller (2011). "Cross-Cultural Cognitive Interviewing Seeking Comparability and Enhancing Understanding." Field Methods **23**(4): 331-341.

Xiao, Y., F. Hou, et al. (2010). "An investigation into socio-economic impact of adverse drug reactions of antibacterial agent irrational use." Chinese Health Economics **29**(5): 94-96.

Zeng, C.-x. and W.-q. Cai (2011). "Analysis of 164 Case Reports of Adverse Drug Reactions in Children in Our Hospital." China Pharmacy **30**: 038.

Zhang, R., K. Eggleston, et al. (2006). "Antibiotic resistance as a global threat: evidence from China, Kuwait and the United States." Global Health **2**(6): 1-14.

Zweig, M. H. and G. Campbell (1993). "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine." Clinical chemistry **39**(4): 561-577.