12-2016

WELLBEING AND DATA QUALITY IN THE AMERICAN TIME USE SURVEY (ATUS) FROM A TOTAL SURVEY ERROR PERSPECTIVE

Ana Lucía Córdova Cazar
University of Nebraska - Lincoln, al.cordovacazar@huskers.unl.edu

Follow this and additional works at: http://digitalcommons.unl.edu/sramdiss

Part of the Community-Based Research Commons, Other Social and Behavioral Sciences Commons, Quantitative, Qualitative, Comparative, and Historical Methodologies Commons, and the Social Statistics Commons

Córdova Cazar, Ana Lucía, "WELLBEING AND DATA QUALITY IN THE AMERICAN TIME USE SURVEY (ATUS) FROM A TOTAL SURVEY ERROR PERSPECTIVE" (2016). Survey Research and Methodology program (SRAM) - Dissertations & Theses. 10. http://digitalcommons.unl.edu/sramdiss/10

This Article is brought to you for free and open access by the Survey Research And Methodology Program at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Survey Research and Methodology program (SRAM) - Dissertations & Theses by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
WELLBEING AND DATA QUALITY IN THE AMERICAN TIME USE SURVEY
(ATUS) FROM A TOTAL SURVEY ERROR PERSPECTIVE

by

Ana Lucía Córdova Cazar

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Survey Research and Methodology

Under the Supervision of Professor Robert F. Belli

Lincoln, Nebraska

December, 2016
WELLBEING AND DATA QUALITY IN THE AMERICAN TIME USE SURVEY (ATUS) FROM A TOTAL SURVEY ERROR PERSPECTIVE

Ana Lucía Córdova Cazar, Ph.D.
University of Nebraska, 2016

Advisor: Robert F. Belli

In this dissertation, I seek to develop a tool for the enhancement of time-use and wellbeing measures from a total survey error perspective. In particular, I evaluate the quality of the time use data produced in the American Time Use Survey (ATUS), by exploring its indicators and identifying its main predictors, including interview rapport. Results from these analyses are then used to evaluate the extent to which certain variables correlate, as predicted, with expected levels of wellbeing.

The first specific objective was to investigate the data quality of the 2010 ATUS by constructing a data quality index. In my dissertation, data quality was operationalized as the degree of completeness with which the ATUS diary was completed. The second objective was to examine whether interview rapport predicts data quality. Interview rapport is understood as the conversational interaction quality that contributes to motivate respondents to be more thorough in their reports and to help them access the required information. Finally, the third objective was to assess the predictive power of activity-based wellbeing measures and other variables assumed to affect overall wellbeing, controlling for the impact of data quality in the prediction model.
Two factors of data quality were found through a confirmatory factor analysis model, one related to the degree of motivation to report and the second one related to memory processes that impact the level of accuracy and detail of activity reports. Gender, age, and education are significant predictors of both factors. Through a structural model, it was also found that interview rapport predicts motivation and memory though in opposite directions: While rapport appears to benefit the motivation to respond, it can be detrimental to remembering details. Finally, when predicting overall health (taken as a proxy for wellbeing), it was found that only when controlling for the memory-related data quality factor was there a relative increase in the amount of variance explained, although it was not a practically significant increase. Further research would be helpful in validating the measurement of data quality and rapport constructs, as well as to more efficiently incorporate data quality and rapport in the prediction of wellbeing.
ACKNOWLEDGEMENTS

My days in Nebraska have been filled with joy and incalculable blessings. Besides the opportunity Allan McCutcheon gave to me to come to study at UNL, I have found wonderful new people whom I am now fortunate to call friends. Along the way, a precious child was given unto me by Heaven. Francisco Antonio, my son, has made me appreciate even more the value of time and to try to be better at differentiating what is true, beautiful, and good from what is not.

First of all, I wish to acknowledge my most graceful dissertation committee, composed of extremely well-versed and rigorous scholars, though their most important characteristic is that they are among the kindest persons I’ve ever met.

I thank, from the bottom of my heart, my advisor Bob Belli and his wife Jessie, who have become like family members to me and my family. I could not have asked for a better advisor. In the PhD students’ world, rumor has it that there exist advisors who literally become a nightmare for their advisees and an inspiration for the PhD comics’ series. In my case, if anything, this was the other way round. Dr. Belli has been a real mentor -- he always treated me as a junior colleague, and consistently trusted in that I would be able to pull it off. I was always amazed at the quickness and thoroughness at which he revised all my material. He has been rigorous in his dissertation advice, but always kind in the way he pointed out any weaknesses. In his office he has a sign that reads: “There is nothing so strong as gentleness and nothing so gentle as real strength”. He really walks that talk.
Dr. Jim Bovaird has been incredibly helpful and patient with me. With him I also realized that not only was I before an extraordinary academic, but also a good family man, whose eyes light up when talking about his daughters. I believe he truly lives by his conviction that achieving a career should enable a better family life and not the other way around.

I consider Dr. Ralph de Ayala to be among the professors who most effectively taught me the building blocks of what one needs to know about statistics. He has also been extremely patient with my constant deadline changes, and has permanently showed me his support with a serious face that ends in a smile.

Larry Williams, our new SRAM director, has also been extremely supportive. After Allan’s death, he immediately stepped in and agreed to be part of my committee. I am fortunate to see his commitment towards our program, and I trust he will make it flourish in innumerable ways.

I am also extremely thankful to Lesa Hoffman for the way she taught me advanced statistics. I still have a long way to go to master those techniques, but she inspired me to want to learn more and teach others.

Barb Rolfes and Renae Reis in SRAM have been a precious gift for me. They have cared for me in so many ways, and I will always be grateful for their smiles and companionship. I will miss both of them tremendously, but I know our friendship is everlasting. I also want to thank my colleagues at SRAM, especially Jinyoung Lee, Matt Hasting, and Lei Wang. Their friendship and work ethic has been a constant inspiration
for me. Antje Kirchner and René Bautista deserve my gratitude as well for being such
good colleagues. I also want to sincerely thank Eva Bachman in Graduate Studies; she
deserves a special recognition for her smile and for being a true advocate for students.

My gratitude towards Allan McCutcheon cannot be put into words. He was an
extraordinary mentor, teacher, and friend. He parted from us too soon, but his memory
will always accompany me. I pray we will one day meet again.

Finally, but by no means least importantly, I wish to thank my family for their
constant encouragement and love. I thank my parents for all the sacrifices they made
during the writing of this dissertation. In order to keep a two-doctoral-student home
functioning, they left the comforts of their home in Quito and came to take care of us. In
the most Christ-like manner I’ve seen, they did our laundry, cooked for us, and gave us
all of their love in the process. They have never ceased to exemplify God’s love for me. I
couldn’t have made it without them. I also deeply thank Paco for believing in me in spite
of all the struggles, but most especially for being a great dad and having played tirelessly
with Francisco Antonio during those long hours when I couldn’t be present.

In memory of Allan L. McCutcheon.
Dedicated to the Lord of Life and Time, and to those He has given me to accompany me through life’s journey: Francisco Antonio, Paco, Guillermo, and Martha.
This material is based upon work supported by the National Science Foundation under Grant No. SES - 1132015. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.
Why study time?

“The humans live in time but our Enemy [God] destines them to eternity. He therefore, I believe, wants them to attend chiefly to two things, to eternity itself, and to that point of time which they call the Present. For the Present is the point at which time touches eternity. Of the present moment, and of it only, humans have an experience analogous to the experience which our Enemy has of reality as a whole; in it alone freedom and actuality are offered to them.”

— C.S. Lewis (The Screwtape Letters)
Table of Contents

Chapter 1—Introduction ........................................................................................................ 1

Measuring Wellbeing by Looking at the Use of Time ......................................................... 3

Time Use Data Quality and its Relationship with Measures of Wellbeing................................. 5

Error in Time Use Surveys ........................................................................................................ 5

Paradata to Assess Interview Rapport ...................................................................................... 6

Research Objectives .................................................................................................................. 7

Chapter 2—Literature Review: Time Use Research and Wellbeing ........................................ 14

Time Use Research ..................................................................................................................... 14

A Brief Historical Account of Time Use Research ................................................................... 15

A “New” Paradigm to Assess Wellbeing .................................................................................... 17

Measuring the Allocation of Time in the 1980’s ...................................................................... 20

Formal System Elements: Resources, Activities (Time) and Outcomes .................................... 25

Moving Forward: The Currency of Life of the National Time Account Framework ................ 28

Wellbeing and Surveys: Calendar and Time Diaries as the Tools to Measure Wellbeing ........ 33

Applying Time Diary and Calendar Methods to the Analysis of Wellbeing .............................. 34

The Use of Time Diaries and Calendars in the Study of Health as the Main Indicator of Wellbeing .......................................................................................................................... 36

Chapter 3—Literature Review: Surveys, Errors, and Quality .................................................. 38

The Total Survey Error Framework and Survey Data Quality .................................................. 39

Limitations of the Total Survey Error Framework .................................................................... 52

The Management Perspective in Quality .................................................................................... 54
Quality as a Third-Level Concept and the Process Perspective ............... 56
The Introduction of Paradata (Process Data) in the Evaluation of Quality.............................................................................................................. 58
The Term ‘Paradata’ and Their Use to Evaluate Data Quality............... 61
Performance and Production Measures .............................................. 63
Audit Trails .......................................................................................... 64
Time Measures for Response Latency Measures ................................ 65
Data Quality in Calendar and Time Diary Methods .............................. 66
Measurement Error ............................................................................ 66
Measurement Error in Surveys from a Psychometrics Perspective ............. 69
Classical true score model (Model 0) .................................................. 70
Platonic scores model (Model for acknowledging the possibility of bias) ................................................................................................. 74
Reliability and Validity ....................................................................... 76
Assessing Measurement Error in Surveys using Reliability and Validity Measures ...................................................................................... 79
Beyond Classical Test Theory: “New” Methods to Establish Validity and Reliability .................................................................................. 81
Nonresponse Error in Surveys .............................................................. 83
Influences on the Decision to Respond to a Survey Question ................. 84
Nonresponse Error in the Time Diary Methods .................................... 85
The Reduction of Error in Timeline Surveys ........................................ 86
The Role of Flexible Conversation Techniques in Reducing Error in Timeline Surveys ............................................................................. 86
The Role of (Autobiographical) Memory in Reducing Error in Timeline Surveys ...................................................................................... 87
Measurement Model for Data Quality .......................................................... 131
Measurement Model for Interview Rapport .................................................. 142
Two-step Approach for SEM ......................................................................... 150
Predicting Data Quality from Interview Rapport ............................................. 155
Inclusion of Control Variables in the Prediction of Motivation, Memory, and Interview Rapport ................................................................. 160
Fidelity Analysis ............................................................................................. 165
Chapter 7—Conclusion and Discussion .......................................................... 170
Summary of Findings ....................................................................................... 170
Indexes of ‘Data Quality’ and Rapport ............................................................ 170
A Brief Discussion of Reflective Vs. Formative Models .................................... 171
Prediction of motivation and memory from interview rapport ......................... 173
Predictors of motivation, memory, and rapport ............................................... 174
Fidelity Analysis ............................................................................................. 175
Limitations, Implications, and Future Directions for this Dissertation .......... 176
Implications for Survey Research ..................................................................... 174
References ....................................................................................................... 180
Appendices ....................................................................................................... 192
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 5.1</td>
<td>Descriptive Statistics of Measures</td>
<td>111</td>
</tr>
<tr>
<td>Table 5.2</td>
<td>Intraclass Correlation Coefficients for Observed Variables</td>
<td>113</td>
</tr>
<tr>
<td>Table 5.3</td>
<td>Cronbach Coefficient Alpha for Data Quality Indicators</td>
<td>120</td>
</tr>
<tr>
<td>Table 5.4</td>
<td>Pearson Correlation Coefficients for Data Quality Indicators</td>
<td>121</td>
</tr>
<tr>
<td>Table 5.5</td>
<td>Cronbach Coefficient Alpha for Interview Rapport Indicators</td>
<td>125</td>
</tr>
<tr>
<td>Table 5.6</td>
<td>Pearson Correlation Coefficients for Interview Rapport Indicators</td>
<td>125</td>
</tr>
<tr>
<td>Table 6.1</td>
<td>Assessment of Model Fit using WLSMV for Data Quality Measurement Model</td>
<td>137</td>
</tr>
<tr>
<td>Table 6.2</td>
<td>Nested Model Comparisons for Data Quality Measurement Model</td>
<td>137</td>
</tr>
<tr>
<td>Table 6.3</td>
<td>Weighted Least Squares Estimates of Factor Loadings for a</td>
<td>138</td>
</tr>
<tr>
<td></td>
<td>Data Quality Measurement Model</td>
<td></td>
</tr>
<tr>
<td>Table 6.4</td>
<td>Assessment of Model Fit Using MLR for Interview Rapport Measurement Model</td>
<td>148</td>
</tr>
<tr>
<td>Table 6.5</td>
<td>Nested Model Comparisons for Interview Rapport Measurement Model</td>
<td>137</td>
</tr>
<tr>
<td>Table 6.6</td>
<td>Robust Maximum Likelihood Estimates of Factor Loadings for Interview</td>
<td>149</td>
</tr>
<tr>
<td></td>
<td>Rapport Measurement Model</td>
<td></td>
</tr>
<tr>
<td>Table 6.7</td>
<td>Assessment of Model Fit using WLSMV for Three-Factor Measurement Model</td>
<td>153</td>
</tr>
<tr>
<td>Table 6.8</td>
<td>Nested Model Comparisons for Interview Rapport Measurement Model for Three-</td>
<td>154</td>
</tr>
<tr>
<td></td>
<td>Factor Model</td>
<td></td>
</tr>
<tr>
<td>Table 6.9</td>
<td>Weighted Least Squares Estimates of Factor Loadings for a</td>
<td>154</td>
</tr>
<tr>
<td></td>
<td>Three-Factor Measurement Model</td>
<td></td>
</tr>
<tr>
<td>Table 6.10</td>
<td>Weighted Least Squares Estimates of Factor Loadings for a Structural</td>
<td>157</td>
</tr>
<tr>
<td></td>
<td>Equation Model predicting Motivation and Memory from Interview Rapport</td>
<td></td>
</tr>
<tr>
<td>Table</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>-------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>6.11</td>
<td>Regression Estimates of the Prediction of Motivation, Memory, and Rapport</td>
<td>163</td>
</tr>
<tr>
<td>6.12</td>
<td>Regression Model Comparing Models using Different Factors as Weights</td>
<td>168</td>
</tr>
<tr>
<td>6.13</td>
<td>Regression Model Comparing Models using Different Factors as Weights Controlling for Education</td>
<td>169</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>5.1</td>
<td>Screenshot of CATI Instrument with Example Responses</td>
<td></td>
</tr>
<tr>
<td>5.2</td>
<td>Conceptual Drawing of SEM Model to be Estimated</td>
<td></td>
</tr>
<tr>
<td>6.1</td>
<td>Distribution of Total Number of Activities Reported in the Diary</td>
<td></td>
</tr>
<tr>
<td>6.2</td>
<td>Distribution of Whether at Least One Basic Activity was Missing in the Diary</td>
<td></td>
</tr>
<tr>
<td>6.3</td>
<td>Distribution of Whether at Least One ‘Who’ was Missing in the Diary</td>
<td></td>
</tr>
<tr>
<td>6.4</td>
<td>Distribution of Whether at Least One ‘Where’ was Missing in the Diary</td>
<td></td>
</tr>
<tr>
<td>6.5</td>
<td>Distribution of Interview Duration in Minutes</td>
<td></td>
</tr>
<tr>
<td>6.6</td>
<td>Distribution of the Presence of at Least One ATUS Error in the Diary</td>
<td></td>
</tr>
<tr>
<td>6.7</td>
<td>Distribution of Rounding Rate</td>
<td></td>
</tr>
<tr>
<td>6.8</td>
<td>Distribution of Number of Consistent (“Reliable”) Items in from Comparison between Diary and Wellbeing Module</td>
<td></td>
</tr>
<tr>
<td>6.9</td>
<td>Distribution of Total Number of Entries</td>
<td></td>
</tr>
<tr>
<td>6.10</td>
<td>Distribution of Total Number of Secondary Activities Reported</td>
<td></td>
</tr>
<tr>
<td>6.11</td>
<td>Distribution of Total Number of Changes in Reported Activities</td>
<td></td>
</tr>
<tr>
<td>6.12</td>
<td>Distribution of Total Number of ‘Where’ Changes</td>
<td></td>
</tr>
<tr>
<td>6.13</td>
<td>Distribution of Total Number of Verbatim Reports</td>
<td></td>
</tr>
<tr>
<td>6.14</td>
<td>Diagram of the Estimated Structural Equation Model Predicting Data Quality from Interview Rapport</td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

One key concern for philosophers and behavioral, social, and health scientists is to assess people’s wellbeing. For instance, Adam Smith\(^1\) was fundamentally concerned with how individuals and societies are able to secure “all the necessaries and conveniences of life,” that would lead to their ‘prosperity’ or ‘happiness’. Within the field of modern economics - arguably one of the most influential disciplines within the behavioral and social sciences - national income statistics such as the Gross National Product (GNP) and the Gross Domestic Product (GDP) were introduced as a means to describe the state of the economy of a country and, for decades, have been looked as reliable measures of economic success or failure. Such measures that were aimed at assessing the *market activity* of a country have increasingly been thought of as measures of societal wellbeing (Stiglitz, Sen, & Fitoussi, 2010). During the 1980s, a growing number of voices (economists and social scientists alike) have expressed discontent with the notion that an increase in GDP could be equated with an increase in wellbeing (Juster & Stafford, 1985b). The creation of a Commission to “measure economic performance and social progress” in 2008, spearheaded by the economist Nobel laureates Joseph E. Stiglitz and Amartya Sen, has been a clear sign of such a dissatisfaction. The report by the Commission was presented in a highly visible international event in September, 2009. The publication that resulted from this Commission’s report was wittily entitled: *Mis-measuring our Lives: Why GDP doesn’t add up*. Its main recommendations included

---

\(^{1}\) An Inquiry into the Nature and Causes of the Wealth of Nations (1776) and The Theory of Moral Sentiments (1759)
emphasizing the household perspective and broadening income measures to non-market activities. For both of these recommendations, the Commission’s authors acknowledge that having information on how people spend their time - comparable both over the years and across countries - is critical, and that more systematic work in the area of time use needs to be undertaken (Stiglitz et al., 2010).

Although the Commission’s creation was a highly visible initiative within the international political arena, the topic was not new. Indeed, the idea of devising an augmented system of economic and social accounts that would recognize the nonmarket activities of households had already been proposed decades before (Juster, 1985b). Understanding how people spend their time is at the heart of such an endeavor. In any case, a new paradigm has emerged, one where it is accepted - at the national and international political levels - that measurements of wellbeing so far have been flawed, and that alternative ones, especially those which involve measures of time use, need to be further developed and applied. The time to accept subjective measurements as legitimate tools to assess wellbeing has finally arrived: only these alternative measurements seem to have the capacity to capture the fact that wellbeing is a multidimensional concept.

In my dissertation, I explore the possibility to enhance the reliability and accuracy of time-use data, and to better understand their relationship with measures of wellbeing. I argue that time use research can provide hard and replicable data that reflect people’s decisions, preferences, attitudes, and environmental factors (Pentland, Harvey, Lawton, & McColl, 1999), with which the many aspects of wellbeing can be assessed. Thus, I will explore the quality of the time use data produced in the American Time Use Survey
(ATUS) by looking at survey quality indicators on the one hand, and their association with the ATUS wellbeing measurements on the other.

**Measuring Wellbeing by Looking at the Use of Time**

Constructing a system with which to measure wellbeing entails understanding how people use their time as well as how previous life events may predict present and future life developments. For instance, evidence of the future detrimental consequences of tobacco use can be better supported with studies that look at the health status of smokers and non-smokers who have been followed throughout a long period of their lives. Thus, time use researchers need to have accurate information about the past and the present, that is, *valid* retrospective and behavioral reports from respondents that can generalize from samples to the wider population. This ambitious goal depends on statistical based analysis for researchers to be able to draw valid quantitative inferences from representative samples of the population. Indeed, from the mid-sixties, (survey) researchers were already seriously studying subjective wellbeing through the use of population surveys², with which they wanted to understand people’s subjective satisfaction with different life domains. One of their main points is that material or “objective” conditions are but “intermediate” outputs, while subjective measures are the “ultimate outputs” of interest. These concepts are related with the idea of “process benefits”, which establishes that what ultimately matters when it comes to “quality of

---

² See Campbell, Converse, & Rodgers (1976) and Bradburn (1969). Bradburn, for instance, affirms that his book is an: “attempt to (…) pursue systematic data collection within the framework of a single unifying concept. This concept-psychological wellbeing, or happiness-has been of great concern to men since recorded history began and has been the object of vast amounts of thought and research for centuries (p. v).
life” are not the objective wealth conditions of a person, but rather the degree of enjoyment of the activities of producing and consuming that wealth (Dow & Juster, 1985; Gershuny & Halpin, 1996).

Panels that use standardized methods and which follow a cohort across several years have been regarded as a reliable way of obtaining life course information (Belli & Callegaro, 2009). However, panels are rather costly and onerous endeavors, and therefore other methods have been devised. Time diary and calendar methods have been proposed as a feasible alternatives to traditional ways of survey interviewing, especially regarding different aspects of people’s wellbeing, including their emotional and physical health (Agrawal, Sobell, & Carter Sobell, 2009; Belli, James, Van Hoewyk, & Alcser, 2009; Belli & Callegaro, 2009; Juster & Stafford, 1985a; Martyn, 2009). Moreover, researchers have established that the most reliable and valid way to collect detailed information about daily, weekly, monthly or even yearly activities is by actually shunning traditional stylized questionnaires and turning to the so-called “conversational approach,” specifically through the use of time diary and calendar methods (Belli & Callegaro, 2009; Juster & Stafford, 1985a). Although these methods forego the standardization of question wording and encourage a more flexible interviewing approach, they are nevertheless able to produce reliable and valid responses, while assisting respondents to remember and correctly report the interrelationships among past events. Time use surveys, which include both time diaries and “life histories” or event history calendars (calendars from this point forward), can be applied to a wide range of disciplines in
health and social sciences research, allowing for a more systematic study of the many components of wellbeing.

**Time Use Data Quality and its Relationship with Measures of Wellbeing**

*Error in Time Use Surveys.* Acknowledging the critical role that surveys play in the construction of a system to measure wellbeing, one needs to carefully consider the elements that contribute to ensure that survey results are valid and reliable. More specifically, in order that behavioral, social, and health scientists may reach valid inferences about how the way people use their time or how their previous decisions in life affect their wellbeing, the quality of the data obtained through calendar and time diary surveys needs to be ascertained.

Asking for information about the explicit details of how people live their lives is, by its own nature, not an easy task: many threats to the validity and accuracy of such reports exist, including those stemming from memory difficulties to retrieve the requested information or privacy concerns on the part of the respondents. In terms of data quality, in my dissertation I will mainly focus on two types of non-sampling error to examine data quality in time use and calendar surveys: measurement error and nonresponse error.

Measurement error, which may be attributed to the respondent, the interviewer, the instrument, or the mode of data collection, occurs when the respondent’s answer deviates from a “true” value. In time diary questionnaires, however, ascertaining a true value is particularly difficult, and measurement error analyses in the form or re-interviews or record-checks are virtually impossible. Thus, one needs to look at other possible indicators of measurement error. In the case of the ATUS, indicators of
measurement error include, among others, reporting a low number of activities, rounding instances, and failing to report a basic daily activity (eating, sleeping, grooming).

Additionally, to assess the degree of measurement error, and taking advantage of the ATUS wellbeing module in which respondents were asked again whether somebody was present during a particular activity, I will conduct a consistency check analysis whose results will be part of the measurement error analysis.

The second source of error to be examined in this dissertation in order to ascertain data quality in the ATUS is that of item nonresponse. This error occurs when the cognitive process of the response fails at some point and a response cannot be offered. In the case of time use surveys, nonresponse can occur when the respondent fails to remember an activity or an episode, fails to provide additional details of the activity or the spell, or refuses to provide an answer altogether.

**Paradata to Assess Interview Rapport.** An additional important aspect in the quest for accuracy is that error prevention and quality improvement can only be achieved by carefully considering the processes that generate the data (Biemer & Lyberg, 2003). For instance, in the case of interviewer-administered surveys, the interaction between the interviewer and the computer, as well as between the respondent and the interviewer, need to be incorporated into the evaluation of survey error. In that sense, apart from the survey answers *per se*, paradata, data about the data collection process, have the potential to provide insights on those interviewing dynamics that most impact data quality. Accordingly, technological advances in the field of survey methodology such as computer assisted interviewing (CAI) methods that are able to record each and every
action of the interviewer, the respondent or even the instrument that is being utilized, have made it possible to obtain such information. It has been argued that these paradata may be useful to reflect the different stages of the cognitive response process when providing a survey answer. The use of paradata when examining data quality is based on the notion that non-sampling error occurs when there is a breakdown in the cognitive response process (Olson & Parkhurst, 2013), and the fact that paradata is deemed capable of capturing such breakdowns. For instance, paradata can help evaluate the processes that impact data quality by providing additional insights of those covert communicative and cognitive processes that occur during the interviewer-respondent interaction (Belli, Bilgen, & Al Baghal, 2013). One behavior that could contribute to reveal those is that related to interview rapport, which in this dissertation is understood as the effort interviewers and respondents make to create a friendly environment for the interview. Several studies have tried to tap into the construct of interview rapport in time diary and calendar questionnaires using paradata (especially through the use of behavior codes) as a means to understand the extent of its impact of data quality. (Belli et al., 2004; Belli et al., 2013; V. A. Freedman, Stafford, Conrad, Schwarz, & Cornman, 2012; V. A. Freedman, Broome, Conrad, & Cornman, 2013).

**Research Objectives**

Given that the approach used in my dissertation is based on the notion that time use surveys are the tools *par excellence* to measure wellbeing, in this dissertation I attempt to assess the extent to which data quality affects the predictive power of those variables that most influence people’s level of wellbeing. Having a sense of the degree to
which inferences about wellbeing are affected by the quality of the collected data is of utmost importance in as much as these inferences may actually influence public policy and, therefore, affect real people’s lives.

In light of the above, my dissertation research seeks to enhance wellbeing data quality assessment through the examination of measurement and nonresponse error in the ATUS, a national probability survey that examines how Americans use their time. In particular, I evaluate the quality of the time use data produced in the ATUS, by exploring its indicators and trying to identify its main predictors, and through those results evaluate the extent to which certain variables correlate, as predicted, with expected levels of wellbeing. To accomplish these aims, I will be analyzing both information derived from substantive answers and paradata in the form of keystrokes or audit trails. Importantly, to examine these relationships, I will be using a statistical technique that has the capacity to take account of measurement error, namely, structural equation modeling, which from my review of the literature, has not been used so far in the evaluation of total survey error.

This dissertation has three main specific objectives:

1. Investigate the data quality of the 2010 the ATUS from a total survey error perspective, by constructing a data quality index that will provide an assessment of the most significant factors that affect the quality of the data. 

   It is expected that those interviews with higher data quality, that is, interviews with a higher degree of completeness, will be indicated by interviews that resulted in a higher number of reported activities, less instances of rounding, fewer number of missing reports of basic daily activities (e.g., eating,
sleeping), and fewer nonresponse items compared to interviews of lower data quality.

2. The second objective is to examine whether interview rapport predicts data quality. For this second objective I will also construct an interview rapport index, in which the observed indicators measuring the rapport construct mainly consist of paradata variables that aim at reflecting the interviewing process itself.

2. *It is expected that a higher interview rapport will contribute to reduce error in the ATUS by increasing respondent’s motivation to respond and by enabling an interview environment in which interviewers can effectively probe for the required information.*

Both data quality and interview rapport will be treated as continuous latent constructs that will be measured through observed variables that include both survey responses and paradata. Further, both indices will be created using latent trait measurement modeling techniques, specifically confirmatory factor analysis (CFA).

The ‘interview rapport index’ aims at capturing the conversational interaction quality that contributes to motivate be more thorough in their reports and help them access information in their memories in a more efficient manner. Indeed, this index seeks to investigate the extent to which a productive interpersonal atmosphere was created during the survey interaction. The observed variables hypothesized to be indicators of an ‘interview rapport index’ include: the number of changes in the *activity* and the *where* reports during the course of the interview, as an indication of a flexible conversation
where “repairs” could be made; the number of secondary activities that were reported by the respondent, as a sign of a more engaging and productive conversation (taking into account that secondary activity reports are not required to be probed); and the number of verbatim reports (as opposed to reports recoded through the pre-established ATUS activity codes) as an indicator of the interviewer wanting to maintain the positive reciprocal environment and flow of the conversation, by not disengaging himself from the conversation, but rather directly typing in the information being conveyed by the respondent.

The data quality indicators are, in reality, indicators of the degree of thoroughness with which the ATUS diary was completed. In that understanding, these include: the number of activities the respondent was willing to report (the larger the number, the more complete the report, as it can be the case that among two respondents with the same number of activities, one decided to report all of them, while another, only the minimum possible number to get through the interview); the number of instances of ‘rounding’ (the lower, the more complete the diary report); the number non-response items (the lower, the more complete the diary), and the (non) failure to report a basic daily activity (i.e., eating, grooming, or sleeping). An additional indicator of the thoroughness with which the diary was conducted is a reliability measure created by looking at the consistency of the answers to a question about whether somebody else was present during the activity, which was asked twice during the interview (one during the diary and the other in the ATUS wellbeing module).
3. Assess the predictive power of activity-based wellbeing measures and other variables assumed to affect overall wellbeing (e.g., income, age, marital and employment status), while controlling for data quality and interview rapport. This objective is based on the notion that the most appropriate way to assess wellbeing is by understanding how people use and evaluate the quality of their time. It follows that in order for time diary data to produce reliable and valid measures of wellbeing, these data need to be of good quality. For this third objective I will use the ATUS wellbeing supplement. However, because of a limitation in the 2010 ATUS questionnaire, in which a question about overall wellbeing was not asked, overall health will be taken as a proxy of overall wellbeing.

*It is expected that controlling for data quality and interview rapport will lead to a higher explanation of the variance in overall health (taken as a proxy of overall wellbeing) from measures of use and evaluation of time, as well as from other important predictors of wellbeing, in comparison to a model that does not control for data quality.*

Finally, this dissertation has an overarching objective to respond to a call made at least twenty five years ago to more fully coalesce the disciplines of psychometrics and survey methodology (see Groves, 1989 and Groves & Lyberg, 2010). This call resulted from the realization that no survey measurement is free from error, and that one cannot simply conduct analyses and derive conclusions without such an awareness.
Psychometrics provide the theoretical and analytical tools to incorporate error of measurement into the analysis and interpretation of data measured through surveys. Thus, in this dissertation, I aim at answering that call by using psychometric techniques such as latent trait measurement models for the construction of indexes, and the implementation of structural equation modeling in the evaluation of data quality from a total error perspective. To the best of my knowledge, there is no research in which data quality, understood as the degree with which a survey is completed, is looked at as a latent factor that can be measured through indirect indicators of error stemming from different sources in a simultaneous fashion. In fact, the survey literature has consistently investigated sources of error separately and many times using only one or, at most, a couple of indicators of data quality at a time (e.g., item non-response, unit non-response, sampling error). Importantly, although there have been some efforts to use latent measures of error, there is no research in which structural equation modeling is used to measure a latent factor of data quality, which will then be predicted by another latent factor related to the interviewing-interaction dynamics that are hypothesized to be associated with error in surveys. Moreover, a connection between the constructs of data quality and interview rapport and their predictive power on other variables, that are intended to be measured by surveys, is also lacking in the literature. In my dissertation, I aim to assess the quality of time use measures, as it has been established that these measures are the most appropriate way to evaluate wellbeing -- which is one of the main areas the discipline of survey research should strive to understand—.
In sum, this dissertation aims to provide evidence for the need to fully acknowledge the importance of having measurements with good psychometric characteristics (i.e., good score reliability and validity) (Kline, 2011) in the practice of survey research, and especially in surveys of health and wellbeing that could affect public policy decisions. The long-term objectives of this research are to apply this perspective to other surveys, especially time use surveys concerning people’s wellbeing, and to incorporate more (or all) sources of error and non-statistical indicators of quality, or what Horrigan, Phipps, & Fricker (2014) call a “common quality framework”, which includes: “(1) Relevance; (2) Accuracy and reliability – sampling errors; (3) Accuracy and reliability- non-sampling errors; (4) Coverage errors; (5) Measurement errors; (6) Processing errors; (7) Revision errors; (8) Modelling errors; (9) Timeliness; (10) Accessibility; (11) Interpretability; (12) Coherence; (13) Cost; and (14) Credibility, Integrity, and Confidentiality” (p. 330).
 CHAPTER 2  
LITERATURE REVIEW: TIME USE RESEARCH AND WELLBEING

Time Use Research

The ability of time use research to understand human behavior and its intrinsic relationship with individual and social wellbeing has been widely accepted and has garnered the interest of researchers from a broad range of disciplines including economics, gerontology, political science, nursing and medicine, psychiatry, health education and research, sociology, psychology, education, social epidemiology, criminology, demography, social work, and survey methodology (Belli, Stafford, & Alwin, 2009b; Pentland et al., 1999).

As noted above, time use research methods consist of both time diaries and “life histories” or calendars. Both are similar in that they collect timeline data – for diaries the timeline is a 24-hour day; for calendars it can range from months to years or even longer sections of the life course (Belli, Stafford, & Alwin, 2009a). Both diaries and calendars examine the allocation of time into different activities, including paid work, personal care, leisure, childcare, and, increasingly, a wide range of health-related behaviors by the different population groups (e.g., women, the elderly, persons with disabilities), at the daily, weekly, monthly, or yearly levels. During the first half of the 20th century, the majority of time studies were conducted in the Soviet Union, Great Britain, and the United States, with some studies conducted in France, Germany and Japan. Currently, almost every nation in the world conducts time use studies of some sort, suggesting that

---

3 For a detailed historical account of time use studies during the first half of the 20th century see: Pentland et al. (1999).
from early on, researchers interested in examining those key events that govern people’s behavior, health, and social interactions, have drawn on the study of how people use their time (Belli et al., 2009b).

A brief historical account of time use research. Time use research -in the form of diaries- emerged during the second decade of the past century in the context of early studies of the living conditions of the working class. That is, the original time use research studies emerged as a response to industrialization and its ensuing pressures on people’s daily lives. Interestingly, calendars also appeared in connection to social issues, specifically with the objective of investigating migration processes to the United States. The two first published works that gave an account of the daily lives of working-class families were published separately in the same year of 1913 in the United States and the United Kingdom (Bevan, 1913 and Pember-Reeves, 1913 cited in Pentland et al., 1999). These works, however, did not include systematically collected and representative data from diaries. The earliest sophisticated diary study belongs to the Soviet S.G. Strumlin, which was intended for use in governmental planning (Pentland et al., 1999). Several other smaller and isolated efforts were launched in Japan, the United Kingdom, and the United States, until the mid-1960s when Alensander Szlai launched a very ambitious program to systematically obtain time diary data from 13 countries around the world. Even though by the mid-1960’s the official statistical agencies of practically all Western European countries were conducting time use studies, the United States only saw its first official time use survey, the American Time Use Survey (ATUS), in 2003.
The first study where quantitative information coming from calendars was collected, processed, and analyzed dates from 1969, when the Argentinean Jorge Balán collected 1640 life histories of men ages 21-60 in Monterrey, Mexico. Before that, Thomas and Znaniecki (1918) had strongly advocated for an intensive use of life histories, but did not pursue systematic data collection because of insurmountable technical difficulties at that time. For Balán, Browning, Jelin, & Litzler (1969), their ability to take advantage of “the possibilities opened up by large-capacity computers” (which, at that moment, involved the use of punch cards for operating and conducting the statistical analyses) made it possible to systematically analyze a large number of life histories (p. 107). In the same year, researchers from Johns Hopkins University conducted the first calendar survey in the United States, which looked at socio-economic wellbeing in America (Belli & Callegaro, 2009); specifically, its purpose was to empirically examine how social groups and individual households attained social mobility in order to identify alternative intervention directions (Blum, Karweit, & Sørensen, 1969).

Almost two decades elapsed before the next systematic study using calendars was conducted by the demographers Freedman, Thornton, Camburn, Alwin, and Young-DeMarco (1988). They sought to accurately measure the trajectories and event transitions that shape the life course (Belli et al., 2009a), and especially to understand the processes that govern the transition from adolescence into adulthood by looking at the different aspects of the lives of a group of 900 15-year olds over a period of nine years. Freedman and colleagues were able to estimate, through the use of sophisticated statistical methods,
dynamic causal interrelations among several aspects of the life course, through which they concluded that the life course is not a unidimensional series of events unfolding, but a “simultaneous unfolding of many dimensions, all interwoven temporally and causally in complex ways” (p. 38).

Many more studies followed, which have shown that time use research not only contributes to explain people’s current condition, but also the long-term consequences of their daily decisions on their later wellbeing (Belli & Callegaro, 2009). As a result, time use research becomes a means not only to understand individual wellbeing, but also social change.

A “New” Paradigm to Assess Wellbeing

In the 1980s, Juster & Stafford (1985b) presented a ‘time use and wellbeing paradigm’ that devises a comprehensive system of measures with which to assess wellbeing (Juster, Courant, & Dow, 1981; Juster, Courant, & Dow, 1985). This system has been expanded over the years and has laid the foundations for the so-called “social accounting systems”. One of the main aspects of these social accounting systems is that they focus on the nonmarket contribution and behavior of households, which goes beyond the typical scope of economic science where only the activities and behaviors that can be measured in monetary terms are included. As put by Juster (1985a) -almost one quarter of a century before Stiglitz and Sen’s Commission- “part of this thrust is motivated by the desire to provide a more accurate assessment of economic and social welfare. Concentration on the market activities of households misses the great bulk of activities that contribute directly or indirectly to wellbeing. In the broadest representation
of an economic and social system, all activities engaged in by individuals need to be brought formally into the analysis framework, *and an attractive measuring rod for that purpose is the way in which people use time* [emphasis added], supplemented by the social context of time use” (p. 19). The accounting notion therefore needs to be broadened through an alternative conceptual framework with which to obtain a more complete representation of the individual and social wellbeing (Juster et al., 1981; Juster et al., 1985).

However appealing this alternative conceptual framework may be, time use scholars concede that there are serious methodological issues involved in the measurement of time use, and thus agree that unless time use can be measured in a valid and reliable way, such an endeavor will not produce interpretable empirical estimates of time use (Juster, 1985a; Phipps & Vernon, 2009; Robinson, 1985). So, notwithstanding the limitations of the traditional monetary measurement system, it has been proposed that it is precisely by emulating and expanding that same system, that an analytic system that looks at how time is allocated to the production of goods and wellbeing may be devised. What is more, such a new system should serve as a bridge between two important groups of social scientists: economists on the one hand, and sociologists, psychologists, political scientists and health researchers on the other. The first group has been mostly focused on material wellbeing, the latter has mostly focused on a different, though complementary type of wellbeing, as reflected in the variegated “social” indicators they have developed, such as depression scales, subjective wellbeing indexes, or even “freedom” indexes to measure the “health” of a democracy. In sum, economists have associated wellbeing with
the flow of material goods and services, while other social scientists and have sought to measure wellbeing with a wider lens, in which the “social indicators” or “quality of life” have been staple terms.

To summarize, the original system designed to measure material wellbeing, which emerged in the 1930’s -mostly as a result of the Great Depression’s aftermath and the need for a formal system of national accounts- was mainly concerned with faithfully representing all those activities taking place in the market, which included the costs of producing output within the business and public sectors of the economy (Juster et al., 1985). Manifold have been the uses and benefits of this traditional system of national accounts. Indeed, as put by Krueger and his colleagues (Krueger et al., 2009), the development of such a system is possibly the chief contribution and most important achievement of the science of economics in the last century. Through it, countries are able to track their national income and thereby to limit fluctuations, a goal of public policy around the world. Likewise, the system of national accounts is used to “estimate bottlenecks in the economy, to forecast business growth, and to inform government budgeting” (p. 9). The main problem, however, is that it fails to account for all the sources of household and individual wellbeing and can only be taken as partial measures of society’s wellbeing. Significantly, the national accounts system misses all the activities that are not formally included in the market, (such as unpaid cleaning and childcare), which actually produce services that otherwise could be purchased on the market. But most importantly, “the National Accounts do not value social activities, such as interactions between friends or husbands and wives, which have an important effect on
subjective wellbeing” (Krueger, 2009a, p. 10). Because of that, the critiques to the traditional economical model of welfare, where time and leisure were inadequately measured, soon emerged (in the late 1950’s), paving the way to the emergence of diverse types of “social accounting systems”, which have the objective of expanding the boundaries of welfare measurement by going beyond the market “price tag” assignment to activities, and incorporating the household element, which inevitably involves a subjective perspective.

**Measuring the allocation of time in the 1980’s.** The ‘time use and wellbeing paradigm’ presented in 1981 by economists Thomas F. Juster, Paul N. Courant and Greg K. Dow was the first formal conceptual system to measure and analyze wellbeing through the use of time (Juster et al., 1981; Juster et al., 1985). The key idea of their paradigm is that “the ultimate constraints determining the level of individual wellbeing are the availability of human time and the set of factors that determine the effectiveness with which time is used” (Juster & Stafford, 1985a, p. 1). That set of factors does not only include material or intellectual resources, but critically, the *capability* individuals have of enjoying and utilizing those resources throughout their life course, which is shaped by levels of physical and mental health. It is precisely the ‘capability of enjoyment’ that time use research has sought to measure.

This “new” social accounting system linking time-use and wellbeing develops from a long-established conceptual framework in economics that is based on the measurement of tangible resources. Monetary values are assigned to each element of the system and concepts such as wages, prices, profits, and interest rates emerge; such
concepts have helped shape the national accounts systems that virtually every nation in the world has implemented to date. As put by Robbins (1952), “the scope of economic science has sometimes been defined to include all activities and behaviors that could be calibrated by way of the monetary measuring rod, and the limits to the boundaries of economics have been set at the point where monetary transactions could not be found”.

This new time use and wellbeing paradigm, on the contrary, comes from the “developing concern over traditional National Income Accounts, and provides a conceptual framework which broadens the accounting notion to give a more complete picture of the sources of household and individual wellbeing” (Juster & Stafford, 1985a). This new paradigm, which emerged with the objective of bridging “the gap between the way in which economists have thought about material wellbeing and the way other social scientists have thought about social indicators” (p. 2), is rooted in a general critique to the traditional system of economic accounts. Four specific main critiques are delineated below.

The first critique centers on the sole focus on flows of material goods and services, where the wellbeing of individuals and societies is determined by the combination of available goods and leisure. In particular, the traditional economic welfare function -in which leisure and goods are the only elements to be considered- lacks any appreciation for a positive connotation for time that is spent working: time at work is always a “bad” and only leisure time constitutes a “good” (Juster & Stafford, 1985a).
A second critique is that the traditional model accounts for the so-called “value
added” to products at the different market stages (i.e., extraction, manufacture, and
distribution), but expressly excludes any type of value added within the household. For
instance, costs of manufacturing and distributing food are accounted for, but those costs
related to the time spent in preparing (nutritious) food are not considered. Therefore, one
of the main differences between both systems is that the social accounting system focuses
not only on resource inputs, but on the changes in output. An illustration of that is the
mother who spends time preparing a nutritious meal for her family or helping her child
with her homework. Here, not only the food ingredients purchased in the market or the
services provided by the school matter; there is value added and a resulting different
output from the time invested in cooking and parenting. The result of incorporating a
measure of the real output (such as a well-nourished family) allows for distinguishing
between intermediate and final output (Dow & Juster, 1985; Gershuny & Halpin, 1996;
Juster & Stafford, 1991; Krueger, 2009). By avoiding an exclusive focus on resource
inputs, the social accounting framework captures other phenomena that also impact
wellbeing, but which traditional economic systems of welfare conceal. As put by Stiglitz
et al. (2010), “traffic jams may increase GDP as a result of the increased use of gasoline,
but obviously not the quality of life” (p. 3). In sum, in the social accounting framework,
material or “objective” conditions are “intermediate” outputs, while subjective measures
are the “ultimate outputs” of interest.

A third relevant critique raised by scholars other than economists about the
inadequacy of the traditional economic model of welfare, and that Juster et al. make their
own, include the lack of distinction between current and capital accounts, both in the household and the government sectors.

In sum, Juster and his colleagues (1985) find three recurrent themes in the writings of all those who had previously criticized the traditional system of measuring (economic) wellbeing: (a) the need to correctly account for the division of societal effort between current and future benefit flows; (b) the need to emphasize the measurement of outputs rather than inputs be emphasized, even if this entails reducing the amount of information there is with regard to the costs of resource inputs; and (c) the importance of incorporating “a variety of unpriced activities into economic accounting systems in a more systematic way” (p. 115). As such, these criticisms to the conventional national accounting system were not only about its “boundaries” (placed at the doorstep of the household), but also about the limited analytic possibilities of the system, which renders an account structure that is inadequate to measure wellbeing.

Finally, the movement towards a more comprehensive system of accounts emerged from a deeper understanding of the concept of utility⁴, which has provided the intellectual basis to incorporate the role of time in the measurement of wellbeing. From the social accounting perspective, utilities do not just depend on the final product that results from a certain personal activity, but also on the enjoyment of the time spent in that activity. The theoretical implication is that the way time is used –its level of enjoyment– needs to be taken into account (Gershuny & Halpin, 1996; Juster et al., 1981; Krueger, Kahneman, Schkade, Schwarz, & Stone, 2009). Hence, what becomes of interest are the...

---

⁴ Juster (1990) reminds us of the original Benthamite concept of utility as the “cardinally measurable psychological flow of satisfactions attached to goods and services purchased in the market” (p. 156).
so-called “process benefits,” that is, the extent to which a person actually enjoys the activity regardless of its price.

The concept of process benefits finds support in two additional notions previously put forth in the literature: (1) That consumers derive utility from multiple features of purchased goods, rather than readily observable flows of goods (e.g., a car’s comfort and speed, rather than just the car’s model and make) (Lancaster 1966); and (2) That there is a “joint dependence” of utility on both the end results of activities (e.g. a meal that results from the activity of cooking) and the individuals’ preferences among the different activities an individual may choose to do (Wachter, 1975).

Concepts such as quality of life and the creation of social indicators emanate from this perspective, as well as the notion that goods and services are instruments for the (subjective) enjoyment of activities rather than ends in themselves. Indeed, for the critics of the traditional economic accounting systems, the underlying concept is that “‘real’ social indicators ought to be measures of subjective satisfaction with various domains of life”, as this is the only way to actually register the “final output” (Juster et al., 1985, p. 117).

It is within this debate and in the aim of providing a more sophisticated perspective that Juster and his colleagues put forward their conceptual framework for the theory and measurement of wellbeing. Two principles guided this seminal work: (a) that the research on the measurement of wellbeing among economists and other social scientists can and needs to be linked, and (b) that both groups of social scientists have
already made important advancements in their respective fields, that will allow to “bridge the gap between analyses of goods inputs and of ultimate outcomes” (p. 117).

**Formal system elements: Resources, activities (time) and outcomes.** Given that the emphasis is placed on the choices made by individuals and households, the system developed by Juster et al. (1985) looks at the relationship of resources, activities (or rather, the time spent on them), and outcomes within households. As put in their own words: “(…) we are proposing a system of social accounts that can best be understood in terms of the way it treats resource constraints and the linkages between these constraints and the generation of preferred outcomes. We view individual choice as constrained by two fundamental factors: by a finite amount of time to allocate among alternative activities, and by a given set of ‘stocks’ or ‘states of the world’ inherited from the past” (p. 122).

Such a system requires to specify the variables to be measured for it to become a truly analyzable system. These variables include:

1. Goods produced in the market (GNP-type)
2. Time allocated to all activities among the population
3. Capital stocks (which range from tangible capital assets such as machinery and houses to intangible assets such as skills, knowledge, health, and quality of the environment).
4. A set of contexts (organizational and sociopolitical) within which activities take place and that are non-measurable counterparts to capital stocks. These
may include, for instance, institutional arrangements, marital status, or geographic location.

5. Outputs stemming from the household production process, including the number and quality of children, meals, and living quarters

6. Objective indicators of societal conditions, indexed by measures such as disability-free days, amount of leisure time, proportion of households with incomes (current and prospective) above a minimum standard, etc.

7. Subjective measures of satisfaction connected to the various conditions of individuals (satisfaction with job, marriage, income, etc.)


Under this framework, “the basic resources available to individuals or society for the production of wellbeing can be defined as total available time, on the one hand, and the stock of inherited ‘wealth’, on the other, with wealth being defined very broadly to include not only conventional capital assets like factories, houses and cars, but also human capital skills, environmental assets, stocks of associations between individuals, and the social and political superstructure. In brief, (...) individual or societal wellbeing [results] from the application of those resources to the production of market and nonmarket output. These outputs are then combined with nonmarket time to produce other nonmarket outputs, and ultimately satisfactions or utilities” (Dow & Juster, 1985), p. 399. It also becomes apparent that the implementation of such a system entails a major task, fundamentally because of the difficulty to reliably and validly measure each of the
variables included in the model. In fact, the first attempt to construct measures of process benefits was not successful, as measures were fraught with validity issues and estimates were not robust when used in different analytical models (Dow & Juster, 1985). The main problem stemmed from two issues: First, the fact that the so-called ‘preference data’ used to measure process benefits were obtained by providing respondents with a stylized list of 22 activities, where respondents needed to rank, on a scale of 0 to 10, their preferences regarding each of the activities, in general, and not the particular activity experienced on a particular moment of time. The activities that were measured included cleaning the house, watching television, reading to children, active sports, among others. The main problem with this approach was that respondents needed to rank their preference of activities out of a general idea of what that activity entailed, introducing social desirability bias in their responses. For instance, when a mother is asked about how much she enjoys reading to her child or cooking a healthy meal for her family, she will base her response on her general beliefs and how much she believes she should be enjoying an activity, and not necessarily on how she felt during the actual experience. Likewise, bias was introduced in this literature as researchers were the ones judging whether an activity was beneficial for the respondents’ welfare (both in the present and in the future). Therefore, some of the data used did not reflect the judgments of the respondents, but those of the researchers (Juster, 1985b).

Such a failure in the empirical attempt of measuring wellbeing is, however, just a superficial one. In fact, the real objective was to devise a formal conceptual model for social accounting, and that was certainly achieved. Noting the difficulties and limitations
of such an endeavor only adds to the theoretical success, for the cornerstone of a new integral system to measure wellbeing had been laid. Furthermore, their work decisively helped to establish time diaries as the most advantageous method to measure time use (Juster, 1985a). In effect, although Juster and his colleagues did not make a direct use of time diaries when evaluating process benefits and wellbeing, they did delve into the conceptual and methodological issues involved in the measurement of time use, as they were convinced that in order to provide a more accurate assessment of economic and social welfare, “all activities engaged in by individuals need to be brought formally into the analysis framework, and an attractive measuring rod for that purpose is the way in which people use time, supplemented by the social context of time use” (p. 19). Furthermore, Juster (1985a) stated that “For purposes of both social accounting and behavior modeling, a uniform ‘currency’ in which concepts can be structured and behavioral parameters estimated is of enormous value. Historically, the only social science with such a currency has been economics, where money has served as a measuring rod by which a large number of decisions can be understood, evaluated, and aggregated” (p. 20). In that framework, Juster posits that certain type of social phenomena may be measured with a ‘time currency’, being time diaries the most accurate method of obtain data on time allocation. Through time diaries, researchers are able to obtain estimates of the time devoted to the different activities, which are reported in the shape of diaries shortly after the event.

**Moving forward: The currency of life of the national time account framework.** Based on the ideas proposed in the seminal work of Juster and colleagues
(Juster et al., 1981; 1985), the pursuit for a new system of accounts that will allow, in practice, to measure wellbeing in a more comprehensive fashion—one which necessarily would include how time is experienced—has gained momentum. During the first decade of the twenty-first century, the idea of establishing a system of “National Time Accounting” (NTA) was presented by a distinguished group of scholars (four psychologists and one economist), with the support of the National Bureau of Economic Research (Krueger, 2009; Krueger et al., 2009). This new system of accounts is dependent on self-reported data of subjective outcomes (or how they call it, “subjective wellbeing”), that is to be measured and reported in tandem with traditional national estimates of a country’s economic activity in order to measure overall wellbeing. In words of the economist of the group, the NTA is a “framework for measuring, comparing, and analyzing the way people spend their time across countries, over historical time, or between groups of people within a country at a given time” (Krueger, 2009a, p. 2).

Significantly, it is based on time use and its affective (emotional) experience. The novelty of this method is that time use evaluation does not rely on researchers’ or coders’ judgments about whether an activity constituted enjoyable leisure or hard or tedious work or home production. In effect, the NTA approach, referred to as the “evaluated time use” approach, relies on individuals’ own evaluations of their emotional experiences during their various uses of times, and it is not up to researchers to group and evaluate the reported activities. This evaluated time use allows respondents to express emotions in a
multidimensional fashion; for instance, they can indicate they were happy, tired, and stressed, all at the same time, while experiencing a certain activity or situation.

Just as their predecessors from the mid-1980s, the proponents of the NTA base their theory on a critique to the National Income Accounts (NIA) system, as such measures (e.g. GDP per capita or consumption per capita) “only represent a component of total welfare because wellbeing depends on more than economic output and material consumption. In addition, aspects of life that contribute to economic output may detract from well-being. For example, an increase in pollution could be associated with decreased welfare but increased production and national income. National Time Accounting partly overlaps with NIA, but also reflects other features of well-being that are not captured by NIA. For example, time spent socializing with friends is not measured in national income but is important for well-being” (p. 2). Similarly to Juster and colleagues, but in a clearer way, Krueger and colleagues (2009) do not wish to substitute the NIA system but to complement it. They acknowledge that the NTA is also incomplete and only provides a partial measure of society’s wellbeing. Still, they argue that the evaluated time use approach is able to provide a valuable indicator of society’s wellbeing, especially because the resulting measures are linked to time allocation. Accordingly, the NTA can offer analytical and political advantages that may not be available from other type of measures of subjective wellbeing including those of overall life satisfaction.

The NTA framework is also built on Juster’s concept of the ‘process benefits’ of activities, defined as “the set of satisfactions generated by activities themselves” that
makes them an essential ingredient of the distribution of wellbeing. However, as already mentioned, Juster did not actually link the satisfaction evaluation to the specific activity reported, but to how individuals enjoyed the activity or situation in general. This approach fails to capture what people actually experience and results in profound discrepancies between concurrent and retrospective reports of affective experience (Schwarz, Kahneman, & Xu, 2009). Therefore, in the NTA framework, time diary methods are utilized in a way that respondents are to connect specific events that actually occurred to the way they experienced them at an affective level. Three potential biases are prevented from occurring: (1) respondents don’t need to develop a theory of how much they should be enjoying that activity in general in order to construct an answer to the question; (2) respondent’s will be less sensitive to the interviewers’ opinion about them, as we are talking only about an specific instance (i.e., respondents will feel less self-conscious by reporting that one particular time they were not happy while taking care of their children, that if they were to say they are not happy about taking care of their children in general); (3) respondents are not put in the position of needing to accurately aggregate their experiences over many times they engaged in a particular activity in order to provide a “general activity judgment”; and (4) the potential for selection bias will be less likely, as respondents will not need to choose from their past memories the best or worst moments of a particular type of activity in which they were engaged (Krueger et al., 2009, p. 15-16).

Additionally, it is important to note that within the NTA approach, the uncertainty of whether individuals interpret enjoyment scales in an interpersonally comparable way is
potentially better handled. Indeed, the NTA framework proposes the U-index, where the U stands for “unpleasant” or “undesirable”, with which the authors try to address the problem of comparability by just focusing on measuring the proportion of time an individual spends in an unpleasant state. This approach allows for the computation of an average U-index for a group of individuals. According to its proponents, “[t]his statistic has the virtue of being immediately understandable, and has other desirable properties as well. Most importantly, the U-index is an ordinal measure at the level of feelings” (p. 18-19).

Two serious criticisms have emerged toward the NTA’s approach to measuring wellbeing. First, the fact that reducing emotional experiences to a dichotomous characterization (pleasant or unpleasant emotions) necessarily reduces the amount of information about the intensity of positive or negative emotions. The second criticism is that the NTA only provides information about episodic feelings and “misses people’s general sense of satisfaction or fulfillment with their lives as a whole” (p. 11). The best example for this is given by Loewenstein (2009) who argues that traveling to a new a different country is fraught with uncomfortable and unpleasant situations, namely airport lines and the flight itself. However, the experiences and lessons gained from traveling abroad may be extremely valuable and, at the end of the day, makes one a happier and wiser person. In spite of these criticisms, Krueger and his colleagues, along with numerous other prominent authors, have been able to provide empirical evidence that self-reported affect, even reduced to a binary scale, can predict important (objective) life outcomes, especially with regards to the quality of individuals’ social life, work stability,
longevity, and the quality of health (Krueger et al., 2009). Significantly, in the areas of health psychology and behavioral medicine, it has been shown that positive and negative affect play a central role in health outcomes, particularly in connection with the translation of the psychosocial environment into physiological states. In sum, collecting emotional experiences directly connected to actual occurrences has proven to be useful in terms of predicting future crucial life outcomes, and thus, there is “signal” in people’s self-reports of their affective experiences which is possible to be analyzed and interpreted. However, given that this potential signal is subject to biases due to the way the data is collected, the method used to do so is of utmost importance, a topic to be expanded upon in the following sections.

**Wellbeing and Surveys: Calendar and Time Diaries as the Tools to Measure Wellbeing**

Constructing a system with which to measure wellbeing entails understanding not only how people use their time but also how previous life events may predict future developments. For instance, evidence of the future detrimental consequences of tobacco use can be better supported with studies that look at the health status of smokers and non-smokers who have been followed throughout a long period of time. Thus, time use researchers need to have accurate information about the past and the present, that is, valid retrospective and behavioral reports from respondents. This ambitious goal cannot be based on qualitative research analysis (e.g. through in-depth interviews) that cannot be generalized. On the contrary, statistical based analysis is needed for researchers to be able to draw quantitative inferences from samples to the general population. For that reason,
interviewing a representative sample of the population is inevitably involved in such an endeavor.

Panels that use standardized methods and which follow a cohort across several years have been regarded as a reliable way of obtaining life course information. However, panels are costly, and thus other methods have been devised. Time diary and calendar methods have been proposed as feasible alternatives to traditional ways of survey interviewing, especially regarding different aspects of people’s wellbeing, including their emotional and physical health (Agrawal, Sobell, & Carter Sobell, 2009; Belli, James, Van Hoewyk, & Alcser, 2009; Belli & Callegaro, 2009; Juster & Stafford, 1985a; Martyn, 2009). Although these methods forego the standardization of question wording and encourage a more flexible interviewing approach, they are able to produce reliable and valid responses. Furthermore, calendar and time diary methods can be applied to a wide range of disciplines in health and social sciences research, allowing for a more systematic study of the many components of wellbeing.

**Applying time diary and calendar methods to the analysis of wellbeing.** An important feature of data collected through a time diary or an calendar is that they can be used not only to create simple individually aggregated summaries, but summaries that can give account of “complex constructions” (Stafford, 2009). For instance, in the area of employment histories, calendars can be used to construct summary employment measures for the entire 52 weeks of the year (such as employment periods, sick periods, being on vacation, and being out of the labor force). In addition, complex constructions can be created for calendars by analyzing descriptors. For instance, by asking about hours
worked in more than one job or about a second job at a lower wage rate, calendars can permit the descriptive characterizations of multiple job holding status over the course of a determined amount of time. Likewise, descriptors can be used to characterize a given spell or episode, as well as to track their sequence. For example, complex constructions can be captured regarding the type of employment spell, such as full-time or part-time, or whether one was unemployed or out of the labor force due to an injury. Descriptors can also tell us whether the weeks of unemployment during a calendar year correspond to one spell or a number of shorter spells, as well as their timing. These distinctions, which are critical for the correct analysis of labor economics, are not available from summary measures using non-calendar methods.

Similarly, the data obtained from individual 24-hour time diaries can be processed into overall time use measures with which to account for how a society as a whole allocates time into the different activities and how these are distributed across different subgroups (e.g., housework for males and females). Complex constructions can also be created through micro-diary activity data, in which descriptors of the activity (where, who with, secondary activities) can be included. Such descriptors are critical to understanding the patterns of social interactions, such as child or elderly care. Similarly to calendars, the resulting complex timelines overcome the limitation of time-use aggregate measures which are not able to track the sequencing of activities. For instance, in a diary using stylized lists, one may know the total amount of hours of television watching or sleep, but their sequencing remains unknown. Finally, diary activity records are also amenable to the inclusion of subjective descriptors such as affect, whereby
elapsed time can be characterized as productive, enjoyable, unpleasant, or meaningful. Including affect descriptors can be extremely useful in the assessment of wellbeing, even in spite of the fact that diaries only provide a small snapshot of the life of respondent (Stafford, 2009).

In summary, time diaries and calendars are a valuable tool for the measurement and analysis of many aspects that impact wellbeing. Their capability of including descriptors reduces the potential for biases that can arise from respondents’ direct reports of activities or spells over a “typical month,” a “typical week,” or a “typical day,” where averages and timing of events can be considerably imprecise due to socially desirable answers or the reliance on stereotypic responses (e.g., 40 hours of work per week). The literature has shown several examples of how time diary and calendar methods can provide a theoretic framework for the analysis of wellbeing through the collection of valid and reliable information.

The use of time diaries and calendars in the study of health as the main indicator of wellbeing. Calendars and time diaries have the capacity to more fully capture concurrent activities or events, as they not only capture incidence of events, but also their timing and patterns (Stafford, 2009; Barber, McNeely, Olsen, Belli, & Doty, 2016). Indeed, calendars and time diaries allow for the examination of timing and sequencing of events in different domains, and thus provide a rich picture of potential causal mechanisms in the development of a person’s wellbeing (Barber et al., 2016). Likewise, timeline interviewing methodology has been shown to produce high-quality retrospective reports. Such positive outcomes result from the ability of calendars to
encourage respondents to reconstruct periods of social (e.g., residence, marriages), economic (e.g., employer names), or health-related (e.g., tobacco use) episodes of activity or statuses, by using chronological time and their previous own experiences memory cues (Belli et al., 2009a). Studies using calendars and time diaries can include a variety of modalities, including face-to-face and telephone modes, paper-and-pencil and computer-assisted interviewing methods, and life course and shorter reference periods (Belli et al., 2001; Belli, Smith, Andreski, & Agrawal, 2007; Yoshihama, 2009).

Because of these advantageous properties, calendars and time diaries have been applied to several areas of research related to the diverse components of wellbeing, such as education, employment, and, more recently, to the consequences of exposure to political violence (McNeely et al., 2015; Barber et al., 2016). Among the various components of wellbeing, using calendars and time diaries to measure health conditions (e.g., women’s sexual health, alcohol abuse, mental health, adolescent health, health status of the elderly) has been dominant. Results seem to point to the fact that health is one of the most important indicators, if not the main indicator of wellbeing. Because of this strong link between health and wellbeing, in this dissertation overall health status will be taken as the principal indicator of overall wellbeing.
CHAPTER 3
LITERATURE REVIEW: SURVEYS, ERRORS, AND QUALITY

1 In those days a decree went out from Caesar Augustus that the whole world [emphasis added] should be enrolled.
2 This was the first enrollment, when Quirinius was governor of Syria.
3 So all went to be enrolled, each to his own town. (Luke 2:1-3)

7 And even the very hairs of your head are all numbered [emphasis added].
So do not be afraid; you are worth more than many sparrows. (Luke 12:7)

As Madansky (1986) has shown, survey methodology has its roots in the old biblical days where the kings of Israel wished to count their peoples. In the New Testament, Luke tells us that these “enrollments” continued under the Roman’s rule. Indeed, according to Luke, a survey (in the shape of a census) was instrumental to the fulfillment of the Messianic prophecy. At the same time, in survey terms, we can take the words of the evangelist to mean that only through divine intervention one could perfectly count things (like the hairs of our heads), as it seems that, since biblical times, discrepancies appeared in the counting of people (Madansky, 1986). It is thus clear that error cannot be avoided in surveys. It was, however, only in 1934 that the concept of survey error was formalized (Neyman, 1934), but just from the perspective of the error that arises from the ability to include everyone (i.e., sampling error) (Biemer, 2010).

Sampling surveys had only been accepted by the International Statistical Institute (ISI) by 1925, when during Institute’s meeting in Rome, a resolution was adopted where the use of both randomized and purposive sampling was accepted. Even then, however, sampling surveys were looked at with skepticism (Lessler & Kalsbeek, 1992). Overcoming such skepticism was largely accomplished by Neyman (1934) who was able to link statistical
theory w to survey methods, paving the way to a new independent scientific field of survey methodology that advanced the rigorous use of sampling techniques. These techniques were further developed and refined by other classical statisticians that followed, such as Cochran and Hansen (Biemer & Lyberg, 2003). In any case, because of the controversies within which sampling theory originated (Lessler & Kalsbeek, 1992), for several decades sampling error was the focus of attention while other “nonsampling” sources of error were disregarded (e.g., errors stemming from respondents, interviewers, or the questionnaire, as well as errors due to nonresponse or coverage). This disregard did not mean that the survey methodologists of the time were lacked awareness that sampling error was but a part of total survey error. Thus, as time passed by and sampling techniques had been even further refined, the notion of a total survey error gradually emerged.

The Total Survey Error Framework and Survey Data Quality

Besides sampling issues, survey practitioners had been aware of other problematic aspects of survey research since the beginning of the 20th century. For instance, A.L. Bowley (another fierce advocate of survey sampling since 1906), in 1926 –one year after the famous ISI resolution accepting sample surveys – highlighted the need to control for the many sources of error (Brewer, 2013; Lessler & Kalsbeek, 1992). In 1944 this need was emphasized by Edwards Deming from the Bureau of the Census who listed the thirteen “factors that affect the usefulness of surveys”. It is here where most of scholars situate the beginnings of the discussion of the “Total Survey Error” framework, the
"central organizing structure of the field of survey methodology" (Groves & Lyberg, 2010). Deming’s (1944) thirteen factors included:

1. Variability in response
2. Differences between different kinds and degrees of canvass (…)
3. Bias and variation arising from the interviewer
4. Bias of the auspices
5. Imperfections in the design of the questionnaire and tabulation plans;
   a. Lack of clarity in definitions; ambiguity; varying meanings of same word to different groups of people; eliciting an answer liable to misinterpretation;
   b. Omitting questions that would be illuminating to the interpretation of other questions;
   c. Emotionally toned words; leading questions; limiting response to a pattern;
   d. Failing to perceive what tabulations would be most significant;
   e. Encouraging nonresponse through formidable appearance;
6. Changes that take place in the universe before tabulations are available
7. Bias arising from nonresponse (including omissions)
8. Bias arising from late reports
9. Bias arising from an unrepresentative selection of date for the survey, or of the period covered
10. Bias arising from an unrepresentative selection of respondents
11. Sampling errors and biases
12. Processing errors (coding, editing, calculating, tabulating, tallying, posting and consolidating)
13. Errors in interpretation
   a. Bias arising from bad curve fitting; wrong weighting; incorrect adjusting; (b) Misunderstanding the questionnaire; failure to take account of the respondents’ difficulties (often through inadequate presentation of data); misunderstanding the method of collection and the nature of the data; (c) Personal bias in interpretation (p. 359).

Just as in Neyman’s case, Deming’s article is a turning point in the survey methods field: it highlights the fact that error in sample surveys goes beyond sampling error and should include all potential sources of error that may arise while planning the survey, collecting and editing the data, and evaluating and communicating results.

Although Deming did not use the nomenclature of total survey error and the concepts
contained in the list were neither exhaustive nor mutually exclusive, one can say he is the one who for the first time outlined the many type of error sources that exist. Deming’s major contribution is this early recognition that for a survey to be “useful”, that is, one where error is contained and quality upheld, a number of different dimensions come into play. A second important contribution of this landmark article is the recognition of a user, as it is implied by the need for a survey to be useful (Groves & Lyberg, 2010). Fast forward to the present days, it is now widely accepted that quality is a multidimensional concept that can take on a number of different definitions. A definition of quality borrowed from the management sciences that is widely accepted is that of “fitness for use”, which in the survey context is translated into the requirement that survey data should be as accurate, timely and accessible as possible, within the cost and time constraints imposed in any type of survey endeavor (Biemer & Lyberg, 2003, p. 13). According to Groves and Lyberg (2010) a third key contribution made by Deming is his recognition of the difference between bias components of error versus variance components of error. Finally, one may argue that Deming’s (indirect) addressing of the topic of quality was visionary in the sense that he was not only focusing on the statistical perspective, he also seems to have put forward a broader notion of quality. For instance, he thought about a topic that would come much later in time, that related to the impact of the process of the data collection on the quality of the data, prefiguring the use of paradata or the data about the process of data collection. In effect, he mentions different types of paradata such as the ‘degrees of canvass’, the timing of the responses, and processing errors, including coding and editing.
Two short years after Deming’s paper, in 1946, the Indian statistician Mahalanobis tried to statistically control for the errors introduced by fieldworkers when collecting agricultural data in his country (Lyberg & Elvers, 2003). This attempt was a very significant breakthrough as it was realized that interviewers, editors, and/or supervisors could produce correlated response variances. Until today the interpenetration method created by Mahalanobis is used in order to estimate correlated variances and effective sample sizes. At the beginning of the 1950s, statisticians from the Census Bureau introduced a mathematical study of what they called “response error”, a concept used to designate “non-sampling errors introduced during the course of data collection” (M. H. Hansen, Hurwitz, Marks, & Mauldin, 1951, p. 147). For them, response errors “may be due to the questionnaire design, the interviewing approach, the characteristics, attitudes, or knowledge of the respondent, or a great many other causes [emphasis added]” (p. 147). The model presented by Hansen and his colleagues is the well-known “U.S. Census Bureau survey model”, which introduced the notion that the total error of an estimate can be measured as the mean squared error of that estimate. This model established for the first time the possibility of formally estimating variance and bias components of the mean squared error; specifically, they were able to show that sampling variance is only one type of error and that survey estimates needed to acknowledge the other sources of error to avoid underestimates of the total error (Biemer & Lyberg, 2003; M. H. Hansen, Hurwitz, Marks, & Mauldin, 1951; M. H. Hansen, Hurwitz, & Bershad, 1960).
Thus, by the 1950s, the most important error sources have been identified and discussed. However, it had taken a great deal of effort on the part of the survey community to universally accept sampling (and its ensuing error) in surveys. Therefore, sampling error was the most treated error from the time Deming’s paper was published until the mid-seventies. By 1974, Dalenius (1974) introduced the term “total survey design” (Groves & Lyberg, 2010), where it was explicitly acknowledged that researchers should be able to control all those error sources that could affect the resulting data, including the design of the survey, the data collection and the evaluation systems. In their article about the past, present, and future of the Total Survey Error framework, Groves and Lyberg (2010) relate that the term Total Survey Error was first introduced in 1979 in a health survey where errors of unit nonresponse, measurement, and processing errors were empirically studied, and where total survey error was formally decomposed for the first time in its variance and bias components (Anderson, Kaspe and Frankel, in Groves and Lyberg, 2010).

Twenty years later an important theoretical advancement was made when Groves (1989) attempted to “consolidate the social science and statistical literatures on survey errors” (p. vii), by initiating an academic debate between the areas of statistics, survey methodology, psychometrics, and econometrics around the topic of survey error. The way that this debate was fostered was by delineating error notions from each of the referred scientific fields within the overarching construct of Mean Squared Error (MSE), which was decomposed in error identified as variance (random) and error identified as bias (systematic). Each of those in turn are broken into errors of observation and errors of
nonobservation. Errors of nonobservation are split into coverage, nonresponse (unit and item), and sampling errors, whereas observational errors (or differences between the reported values to a survey question and the “true” values) are split into errors stemming from the interviewer, the respondent, the instrument, and/or the mode. Accordingly, there are errors of nonobservation and observation that are either variable (variance) or fixed (bias). Such classification is then summarized in the popular formula of MSE:

\[ \text{Mean Squared Error} = \text{Variance} + \text{Bias}^2 \]

Groves’ main objective was to clarify the structure and language of errors used across disciplines, and identify whether and which terms were equivalent. For instance, he concluded that the concept of reliability in psychometrics is equivalent to that of survey response variance in survey methods, whereas the relationship between the concepts of validity and survey error was much more complex, mainly because of the many notions of validity that exist (Groves, 1989; Groves & Lyberg, 2010). Finally, a significant area of intersection between the psychometrics and survey literatures was identified by addressing the topic on how in surveys and in psychological measurement there is a need to measure underlying characteristics that are, by and large, unobservable. Indeed, extending what is known as Classical True Score Theory in Psychometrics, it is argued that it is possible to estimate error properties of survey items. As an extension, structural modeling techniques have also been used to estimate measurement error (Groves, 1989, p. 336). This attempt to bring together psychometric notions of measurement error and survey notions of measurement error was an important step in the evolution of the concept of total survey error, especially because it led to the
development of complex models to estimate measurement error (mainly through estimating interviewer variance and memory errors) (Biemer & Stokes, 1991).

A few years after Groves’ 1989 classic book, a somewhat different approach to address error was taken by Lessler and Kalsbeek (1992). In it, they attempted to continue with a debate in which a unifying taxonomy of error in surveys was being presented, in which error was divided as either as either sampling or nonsampling error. The rationale for this division is that sampling error exists by design and is the result of a deliberate decision to study only a subset rather than the entire population. As it is based in a well-established theory, sampling error control is feasible. Nonsampling error, on the other hand, is the result of “mistakes and deficiencies during the development and execution of the survey procedures”, which can either be considered avoidable or the result of the intentional choice to use a certain method to conduct the survey (Lessler & Kalsbeek, 1992). There are a myriad and intertwined opportunities for nonsampling error and therefore, assessing its impact turn into a very complex endeavor. To this end, these authors considered three sources of nonsampling error: the sampling frame being utilized, the failure to obtain a response from members of the sample (i.e., nonresponse), and imperfections that arise in the data collection process (i.e., measurement error).

Importantly, they explicitly talk about the concept of “total survey design” which is defined as “[t]he attempt to control the total error of estimates considering all sources of error” (Lessler & Kalsbeek, 1992, p. 5). They recognize three stages within the total survey design: the planning, execution (which involves data collection and data processing), and the analysis and reporting phase. The notion of quality is included in this
discussion, although it is only mentioned during the execution stage where “the practice of total survey design involves the use of quality control procedures that monitor progress of the data collection and data processing. The goal of the quality control procedure is to detect errors when they occur or soon after so that the survey work can be repeated if necessary. Response rates, item completion rates, edit failure rates, consistency checks, resurveying and recoding are methods used to detect errors in the ongoing survey process” (p. 6). Very important for the discussion in the present work, the use of paradata is already implicitly mentioned as information of item completion, edits, and recoding – all of which comfortably fit in the concept of paradata to be discussed below.

At the beginning of the 21st century, there was a conceptual change with regards to the error in surveys. Specifically, a more explicit link between error and quality started to emerge. Biemer and Lyberg pioneered this movement in their 2003 volume Introduction to Survey Quality. In this volume, the focus again is between the division between sampling and nonsampling error, though this time the list of nonsampling errors was expanded to five sources: specification error, fame error, nonresponse error, measurement error, and processing error. As before, each of these errors can be decomposed into variance and bias, and their formula again includes the summation of variance and squared bias. However, they go one step forward by trying to identify the risk of variable errors and systematic errors by major error source. For instance, they recognize that sampling error has a low risk of systematic error but a high risk of variable error; likewise, nonresponse error has a low risk of variable error and a high risk of bias,
whereas measurement error has a high risk of both, variable and systematic error. Thus, their formula for MSE tries to be more specific with regards to such risks (p. 59):

\[
MSE = Bias^2 + Variance \\
= (Bias_{Specif} + Bias_{NR} + Bias_{Frame} + Bias_{Measurement} + Bias_{DataProcessing})^2 \\
+ Var_{Sampling} + Var_{Measurement} + Var_{DataProcessing}
\]

This formula has two practical implications: (a) It provides a method of assessing the MSE of an estimate by estimating the eight specific components, and (b) it provides a method of reducing the MSE by devising a survey design that minimizes the contribution of each component to the total MSE (p. 59).

However, the most important contribution of Biemer and Lyberg’s volume is that the notions of process quality and total survey error were jointly presented for the first time (Groves & Lyberg, 2010). In effect, Biemer and Lyberg (2010) argue for the importance of the topic of quality because they believe society has gone through a “quality revolution.” Specifically, borrowing notions from the management literature, they contend that survey organizations should not be seen as different from any other organizations in society regarding their need for continuous improvement and the quest for quality. Although they acknowledge that quality is a vague concept, they concede that one of the most general and widely used definitions is the one put forward by Juran and Gryna in 1980 which simply defines quality as “fitness for use” (Biemer & Lyberg, 2003, p. 13). In the survey context, “this translates to a requirement for survey data to be as accurate as necessary to achieve their intended purposes, be available at the time it is
needed \textit{(timely)}, and be \textit{accessible} to those for whom the survey was conducted. Accuracy, timeliness, and accessibility, then, are three \textit{dimensions} of survey quality” (p.13). In that context, the authors acknowledge that the quality of a statistical product is a multidimensional concept: it involves the referred elements of accuracy, timeliness, and accessibility, but also others such as richness of detail, level of confidentiality protection, relevance, coherence, and comparability, among others. As it will be expounded below, survey organizations (public and private) have found the need to have a much more encompassing working definition of quality, mainly out of the realization that users are not just interested in survey error or the accuracy of the estimates provided.

The attempt to link the total survey error notion to the idea of survey quality continued throughout the years. For instance, in their 2004 volume, Groves and his colleagues (Groves et al., 2009) talk about the “the life cycle survey from a quality perspective”. For them, the concepts of quality reflect “mismatches” between successive steps in the survey process and thus the “quality components” are equated to the notion of error. The steps of the survey process follow two parallel paths that come together to produce one survey statistic. The two parallel paths include: (a) the inferential process that allows to measure a construct through the response to a question, and (b) the inferential process whereby population values are represented from a sample to the target population. The errors across the measurement path include: validity errors, measurement errors, and processing error; the representation path errors include: coverage error, sampling error, nonresponse error, and adjustment error. Two important characteristics of these quality components (i.e., errors) is that: (a) each one has both verbal descriptions
and statistical formulations, and (b) each component is the property of individual survey statistics and not of whole surveys. In sum, in their view, a quality survey is one where the designer is able to minimize error in survey statistics by “making design and estimation choices to minimize the gap between two successive stages of the survey process” (p. 49).

Finally, in one of the most recent discussions on the topic of survey quality Lars Lyberg (2012) still maintains that “[s]urvey quality is a vague, albeit intuitive, concept with many meanings” (Lyberg, 2012, p. 107). In this paper, the author recognizes that the survey quality concept has originated from two different scholarships that have only recently engaged in a dialogue: one is the survey error paradigm and the other the quality management philosophies. In his analysis, Lyberg explores the impact of the latter on statistical organizations and he places an emphasis on the necessity of incorporating the process quality perspective. Lyberg’s discussion on data quality and the total survey centers around the notion that the survey error paradigm rests on four pillars that provide the principles that guide survey design, survey implementation, survey evaluation, and survey data analysis (Lyberg, 2012). These four pillars imply that: “We should design surveys so that the mean squared error of an estimate is minimized given budget and other constraints. It is important to take all known error sources into account, to monitor major error sources during implementation, to periodically evaluate major error sources and combinations of these sources after the survey is completed, and to study the effects of errors on the survey analysis” (p. 107). In the historical account of quality in surveys that he provides, in the beginnings of the discussion about 60 years ago, quality was
mainly understood in terms of quality control of the survey operations. As he acknowledges, researchers were well aware that statistics were plagued by errors other than sampling error, but they still had to develop and refine the idea of process quality in order to find a way to systematically reduce those errors and biases (p. 108). Likewise, during those initial times, Lyberg reaffirms that the user was merely an ‘obscure’ player, and its role was very limited when it came to the discussion of quality or design decisions.

Lyberg (2012) notes that because sampling error had been the main focus for so long, during a very long time data quality simply meant having a small Mean Square Error, that is, for data to be quality data, it needed only to be accurate. In other words, good quality data were those which had a small variance, and the squared bias was sometimes not even looked at. Lyberg (2012) recognizes that only later would the ideas of relevance and timeliness would be included in the discussion of survey quality. Many authors have situated the origin of the quality debate in connection to the total survey error paradigm, which in a nutshell, is “a theoretical framework for optimizing surveys by minimizing the accumulated size of all error sources, given budgetary constraints” (Lyberg, 2012, p. 109). The practical implication of such definition is that survey researchers aim at minimizing the MSE for the selected survey estimates that are of most interest to the main stakeholders (Lyberg, 2012). In theory, the MSE will be able to incorporate all error sources, which in turn would allow for a balance between errors and costs (i.e., the optimal design). However, since early in the process, survey scholars realized that it was very difficult to come up with a survey design formula that could
satisfy such of a need. The estimation of sampling variance had been, for the most part, achieved. However, estimation of bias posed special challenges. The best way survey scholars found to deal with bias estimation was through the comparison of estimates obtained from “regular operations” (large-scale ones) to those deemed as “preferred” (where interviewers, coders, supervisors were under more controlled conditions); this approach is now known as the “gold standard” approach, which in theory, is the only one with which bias may be estimated, but which is still very limited when it comes to assess simultaneously all the sources of variance and bias.

To summarize, along the seven decades that have elapsed since Deming’s 1944 seminal paper, survey methodologists have greatly refined the concept of error in surveys by better discriminating among the many and varied sources or error (e.g. sampling, coverage, nonresponse, measurement, processing). Importantly, researchers have established a sophisticated classification of errors according to different (though interconnected) dimensions, such as errors that can be categorized as variance or bias, errors of observation and nonobservation, and sampling and nonsampling errors (Biemer & Lyberg, 2003; Groves, 1989; Groves & Lyberg, 2010). Thus, the notion that survey quality is just a function of the amount of error in the data is no longer tenable. For instance, even if the data are perfectly accurate, if they come from a too small or unrepresentative sample, then one may attain biased inferences. Quality is therefore a multidimensional concept that includes other dimensions besides accuracy, such as those of “relevance, timeliness, accessibility, comparability, coherence, and completeness” (Biemer & Lyberg, 2003). Accuracy is, however, the primus inter pares -the cornerstone.
of quality-, since “without it, survey data are of little use. If the data are erroneous, it does not help much if relevance, timeliness, accessibility, comparability, coherence, and completeness are sufficient” (Biemer & Lyberg, 2003), p. 24). In fact, even though all the other attributes of quality are important, they are to be viewed more as constraints on the survey process rather than dimensions themselves. This broader approach to survey quality has led to the incorporation of ideas from other sciences, especially those from the management and business, as will be expounded next.

**Limitations of the total survey error framework.** Several limitations can be pointed out in the total survey error framework. The most salient was is the lack of measurements of all the MSE components on a routine basis across institutions and countries. This is easily explained because “there is no measure of total error that would take all error sources into account, either because a lack of proper methodology or that some errors defy expression” (Lyberg, 2012, p. 110). Indeed, Groves and Lyberg (2010) consider that the fact that fuller measurement of the statistical error properties of survey statistics have not been achieved is the “great disappointment” of the framework. They have found very little evidence that in the current practice of surveys anything above sampling variance is measured in a routine way.

Another weakness is that the survey models that were designed decades ago, and that are still in use, are limited to measurement and sampling error. This leads to another problem: the estimation of the variance components of these errors require either interpenetrated designs or re-interviewing. These two methods are very costly and most of the times cannot be afforded.
A third limitation of the method is that the existing models do not explore the causes of error itself, they just are specified as variance component models that do not seek to understand the nature of the cause of error. Groves and Lyberg (2010) causal models of statistical error are preferable because if the causes of error are identified, then error can be predicted more accurately and even eliminated. According to them, the increasing use of paradata is to control processes and to conduct analysis of the problem root-cause is an alternative to more extensive evaluations of mean square error components.

A final weakness of the total survey error framework is that key quality concepts have been excluded from it. For instance, the user perspective of quality has, by and large, been absent (Groves & Lyberg, 2010). Likewise, the study of concepts such as relevance and credibility is very rare and we do not really know what is meant by those terms. The problem with the relevance issue is that “[i]nnaccurate statistics are not relevant, but irrelevant statistics might be accurate” (Groves & Lyberg, 2010, p. 863). Likewise, credibility –understood as trust– is vital for statistical and survey organizations because if their products are not perceived as credible by the user, then it doesn’t matter how accurate or relevant they may be. Therefore, it is important that additional notions of quality be included, as they go beyond the statistical estimate itself. As put by Groves & Lyberg (2010), “timeliness, relevance, credibility, accessibility, coherence, and other terms speak to how a particular use of a survey statistic matches the characteristic of the estimator” (p. 864).
In sum, the lack of a ‘design formula’ for surveys that incorporates all the error sources is still non-existent. Many limitations remain within the total survey error paradigm (Groves & Lyberg, 2010; Lyberg, 2012) such as having the focus of quality of statistical products on variance components, particularly on measurement error variance. Similarly, the user perspective is still deficient, for the majority of users are not given the opportunity to question design options or accuracy measures. Despite these and some other limitations of the total survey error framework, it is still considered a very useful one for the analysis of error and the enhancement of quality. Its strengths include that the framework: is able to provide a decomposition of errors; provides a way to separate (at least conceptually) variance from bias, an important matter as each of them affect statistics in different ways; permits a useful focus on errors of observation versus error of nonobservation; and finally, it provides a conceptual foundation of the field of survey methodology that serves as an organizing principle for the classification of the survey methodology literature (Biemer & Lyberg, 2003; Groves & Lyberg, 2010; Lyberg, 2012).

The Management Perspective in Quality

As reviewed by Lyberg (2012), there are two development paths for the current understanding of survey quality: the total survey error paradigm and the one based on the quality management sciences. Remarkably, in the development of both paths, statistician Edwards Deming (also known as the “Father of Quality”) was involved as well. This second path originated in the need of many statistical organizations of industrialized countries (US, Canada, Australia, Sweden) to demonstrate their usefulness lest their
funds be restricted. These organizations found inspiration in management theories and methods, specifically on the approach of quality management. This literature was focused on studying the role of the customer, leadership matters, and the notion of permanent quality improvement (Lyberg, 2012). Importantly, the work by Deming was influential not only among survey researchers but on management scholars as well. He was convinced that statistics played a fundamental role in quality improvement and greatly refined many ideas such as control charts, process variation and common and special cause variation (Lyberg, 2012).

A number of quality frameworks or models have been devised through the years including the Total Quality Management (TQM) or the standards system put forward by the International Organization for Standardization (ISO) among others. These frameworks are very similar between each other and their main objective is to set a criteria for excellence, which in the area of surveys, have been combined with quality strategies from statistical organizations from the mid 1990’s (Lyberg, 2012). A central idea that was imported from the management sciences into the survey practice is the inclusion of the user’s perspective. Indeed, with the drastic technological changes, the user is in a better position to demand more space in survey decisions. That is why, although it is true that without accuracy, other dimensions are irrelevant, the opposite is also true: “very accurate data can be useless if they are released too late to affect important user decisions or if they are presented in ways that are difficult for the user to access or interpret” (Lyberg, 2012, p. 113).
This interrelationship between dimensions also means that sometimes they can enter into conflicts between each other. For instance, there is a permanent conflict between accuracy and timeliness: it takes much longer to get accurate data. Likewise, there can be a conflict between comparability and accuracy: applying new and more accurate methods can damage comparability possibilities across time. Because of this, nowadays it is the norm that organizations have accepted that quality is a multi-faceted concept that needs to take into account the user perspective, including aspects such as customer satisfaction, communication with customers, process variability, best practices, continuous quality improvements and so on (Lyberg, 2012).

**Quality as a third-level concept and the process perspective.** Just as any other organization wishing to be competitive, those organizations devoted to the production of statistics also understood the need to maintain continuous quality improvement processes. The core aspect of such a process is the possibility to measure any type of changes that can index improvement (or lack thereof). According to Lyberg (2012), the measures that can be used by a statistical organization to improve are the same as those of any other businesses. These measures can be based on the so called business excellence models, which across the board include principles such as results orientation, customer focus, leadership and constancy of purpose, management by process measures and facts, personnel development and involvement, continuous learning, innovation and improvement, development of partnerships, and public responsibility (p. 114). For instance, the European Statistical System has adopted the European Foundation for Quality Management model as a way to achieve organizational quality among national
statistical offices. As survey quality researchers have noted, the motivating idea is that good product quality can only be achieved with good underlying processes, that is good process quality (Biemer & Lyberg, 2003; Lyberg, 2012). This notion of quality results in a three-level concept that includes: product level, process level, and organization level.

The product quality level has to do with the “deliverables” that can be one or more estimates, datasets, analyses, questionnaires and so on. Specifically, measures of accuracy (e.g., the MSE) and margins of error belong to this level. The main ‘stakeholders’ in this level are the user or the client.

The process quality level is related to the notion of quality assurance or quality control, whereby the processes that are involved in producing a deliverable are ascertained so that these processes actually deliver what they are expected to deliver. The final objective of the incorporation of this notion is to build quality into the process through quality assurance. It should be noted, though, that quality control efforts only serve to verify if the process is working as intended – quality control does not, by itself, build quality into the process. The way to measure process quality is through selection observation and the analysis of key process variables, also called paradata. The main stakeholder of this level is the survey designer.

Finally, the organizational quality level is related to quality in its broadest sense. Issues such as leadership, competence development, funding, etc. are elements of this level. Indicators of organizational quality can be obtained through user surveys and staff surveys. The achievement of quality at this level involves following a business model, or
implementing and adapting reviews or audits. The main stakeholder in this case is the organization’s management.

The incorporation of management principles in the survey work highlights the importance of having a process perspective in the production of statistical results. This notion comes from the belief that if the production process is seen as a series of steps geared towards obtaining a specific goal that satisfies the needs of the user, then good product quality will result. In that sense, process quality is “an assessment of how far each step meets defined requirements or specifications” (Lyberg, 2012, p. 118). The way to measure and control process quality is by collecting those key paradata (i.e., process data) that have an important effect on the final result. The way to distinguish those key process variables is by identifying stability and variation in the targeted processes. Useful in this endeavor are methods imported from statistical process control such as the special and common cause variation perspective, and from the quality management sciences such as flowcharts and Pareto diagrams.

**The Introduction of Paradata (Process Data) in the Evaluation of Quality**

The literature on surveys and quality tends to be focused on examining the survey process, where literature from the management sciences, specifically quality management models and business excellence models have become relevant. According to Biemer and Lyberg (2003), this movement is based on the notion that “product quality is achieved through process quality” (p.15), a notion ever increasing in survey organizations worldwide. This “process view of survey work extends to almost all processes in a survey organization because many processes that support survey work have an effect on
the quality of statistics products” (p. 15). Important developments with regards to the integration of management sciences and the total survey error framework have occurred within government statistical organizations. Indeed, a number of them have produced protocols and codes of practice about how to deal with the current quality challenges. These include the Australian Bureau of Statistics, Statistics New Zealand, Statistics Netherlands, Statistics Denmark, Statistics Sweden, US Census Bureau, among others (Biemer & Lyberg, 2003). International organizations such as the OECD and the United Nations have also produced documents about statistical quality.

Edwards Deming’s ideas seem to once be influential in this area as well, specifically in realizing that in order to integrate the concepts of survey measurement and process improvement there is a need to obtain measurements, which, to be useful, should be more than a number – they need to have a context (Dippo, 1997, p. 459). Following Deming’s theoretical approach towards process quality, Dippo (1997) reminds us about Deming’s 14 points on quality, in which the 3rd and the 4th ones talk about the fact that any continuous process improvement will “require statistical evidence that quality is built in” and will “depend on meaningful measures of quality, along with price” (Deming 1982, cited in Dippo 1997, p. 470). Dippo therefore concludes that any process improvement effort depends on having good measures of quality built into the production system. Moreover, it is through Deming’s philosophy, whereby quality management is built on sound statistical and scientific principles, that the connection between survey measurement and process improvement is made evident. Indeed, according to Deming, for any product, if one can define the production process and a measure of quality, one
can implement the philosophy and improve the product. Thus, as put by Dippo (1997):
“A survey measure is an easily identifiable product, several process models exist for
survey measurement and there exists a whole slew of quality measures—of the survey
measure and of the process. Thus, there appears to be a natural link between survey
measurement and process improvement” (p. 470).

A corollary of the latter is that evaluation of quality is “an ongoing part of the
survey process that occurs (or should occur) before, during and after each round of data
collection” (Couper, 1998, p. 41). In this sense, process quality goes beyond the notion
of quality understood as the effort to minimize mean square error for a given cost, but
process quality will influence the other elements of quality such as relevance, timeliness
and cost efficiency (Couper, 1998).

As a result, the quality approach in surveys is linked to an effective control of the
total error by the careful consideration of the survey procedures using what we now
understand as paradata. Lyberg (2012) gives an account of the basic design approach
suggested by classic survey methodologists such as Hansen, Dalenius, and others where
paradata were already envisioned:

- Specification of an ideal survey goal.
- Analysis of the survey situation regarding financial, methodological and
  information resources.
- Developing a small number of alternative designs.
- Evaluating the alternatives by reference to associated preliminary
  assessments of MSE and costs.
- Choosing one of the alternatives or a modification of one of them or
  deciding not to conduct a survey at all.
- Developing the administrative design including feasibility testing, a
  process signal system (currently called paradata), a design document, and
  a Plan B. (p.110)
Likewise, Kish emphasized the importance of having small biases, but he also understood that reducing one bias may lead to an increase in the total error. In his 1965 volume, Kish was very clear about the need to strive for a reasonable balance between the various error sources and how error structures may very under different design options (Kish, 1965). To obtain such a goal, Kish also introduced the notion of what we currently understand as paradata, as he also believed that relevant information should be simultaneously recorded during data collection. In any case, both Kish and Hansen and his colleagues were convinced that the practice whereby sampling error was the only error measured needed to be overcome (Lyberg, 2012, p. 110-111).

The term ‘paradata’ and their use to evaluate data quality. Most textbooks and published articles agree that it was Mick Couper who in 1998 verbally introduced the term paradata when presenting a paper at the Joint Statistical Meetings of the American Statistical Association. In the paper entitled, “Measuring Survey quality in a CASIC environment,” (Couper, 1998) reflected on the changes that had occurred in the last decades in connection with the way survey data were collected, specifically the movement towards the so called Computer Assisted Survey Information Collection (CASIC) that occurred in the 1990s. He recognized that the survey methodology field encountered numerous challenges in adapting to the changes, especially regarding the “human side of the transition” and the “optimal use of the new technologies” (p. 41).

Some of the challenges arose because former distinctions between different phases of the survey have started to disappear, such as with the advent of computer assisted interviewing (CAI), the data collection and the editing/coding phases can occur.
simultaneously. Thus, a connected challenge is the adaptation of any previous continuous quality improvement processes that were in place to this new framework in which the boundaries between phases are blurred. Likewise, any quality assessment processes should now incorporate the new ability CAI methods have of automatically collecting a wealth of real-time data about the progress of a study. It was argued back then - and is still today - that these process data that were automatically collected offered an exceptional tool to optimize quality and minimize costs.

Far from supplanting previously used quality indicators (e.g., measures of unit and item nonresponse or interviewer behavior) the new process data generated by the CAI methods complement and expand the possibilities of evaluating data quality. However, the enormous amount of data automatically generated makes the questions of what is needed to be measured, why, and how, even more relevant, as not all these data will necessarily be useful to for the purpose of assessing survey quality. Another aspect to take into account with the advent of CAI methods is that survey instruments have become dramatically more complex. This is the result of the ability computerized instruments offer to introduce more complicated skip patterns and/or screening questions into the surveys. While it is true that the computer assists the interviewer in the routing of questions, the interaction between the computer and the interviewer becomes more complicated than before when only paper and pencil were involved, precisely because of the new possibilities opened by CAI methods, such as the possibility for researchers to include text with instructions besides the question to be asked, as well as the ability a computerized instrument offers to change answers, review previous items, press “hot
buttons” to request help, or temporarily pause the interview. Since some types of paradata can capture every single key pressed by the interviewer, the possibility to evaluate the usability of survey instruments is now possible. This is a particularly important tool, for previous empirical evidence has found that interviewer difficulties with the computerized instrument can interfere with the interviewer-respondent interaction and have an impact on the quality of the data (Couper, 1998; Couper & Schlegel, 1998; S. Hansen, Couper, & Fuchs, 1998).

Finally, it should be noted that paradata need to be ‘mined’ as the result of their own nature: paradata or process data are not necessarily collected with a defined research objective in mind. Importantly, the largest part of the paradata currently being used to evaluate data quality is generated as a “by-product of computer-assisted data collection” (Kreuter, 2013, p. 2), which by and large entails massive amounts of unstructured data that are increasingly available through computer-administered surveys. Thus, the growing availability and amounts of paradata make it necessary to identify which kinds of paradata are useful in understanding these underlying processes and which are not.

**Performance and production measures.** Thanks to CAI methods, it is possible that completed interviews be transmitted to a head office on real time, and therefore the information can be evaluated instantly or at least on a daily basis. Data, which are already in electronic form, also are tabulated instantaneously and project managers scan monitor the progress of the study as it progresses. Thus, besides the substantive answers, a variety of summary process measures are immediately available from the computerized instrument including observations of item and unit nonresponse rates, response variances,
sampling and coverage errors. Scheuren (2001) calls these “macro-paradata” and understands them as a byproduct of sample selection and survey administration (p. 1).

Performance and production measures also include the information coming from the case management system that gives an account of the status of any case at any time. Number of calls or visits, including the time of those calls, to a particular potential respondent or sample unit belong in this category. These types of paradata can provide very useful information to understand cooperation propensity of respondents and/or interviewers’ capacity to obtain a complete interview.

**Audit trails.** Audit trails, also known as keystroke files or trace files are another automatic byproduct of CAI systems (Couper, 2008). Audit trails can record each and every key or command pressed by interviewers (or respondents in a self-administered questionnaire) as they navigate through the computerized questionnaire. Originally, audit trails were designed to assist programmers in debugging instruments, as they allow a replay of the instrument. Their main advantages are that they are easily available, practically free to collect and are practically non-intrusive – neither the interviewer nor the respondent are aware of their collection. Their disadvantages include that by their own nature, audit trails cannot be collected in a consistent or “rectangular” manner, as each individual survey will have its own particularities depending on the way it developed (e.g., call-back was programmed, refusals, item-nonresponse, different skip patterns, etc.). Another disadvantage is the massive amount of (unstructured) information they produce, which can easily and quickly overwhelm the analyst. In fact, only summarizing them poses a large enough challenge. A third disadvantage according to
some authors is that audit trails only capture the interaction between the computer and the interviewer, and cannot address the interactions between interviewers and respondents (Couper & Schlegel, 1998; Couper, 1998). Because of this limitation, Couper and colleagues have argued that audit trails can only be used as a supplement to other methods of interviewer or instrument evaluation. In my dissertation, however, I present evidence that audit trails can also provide indirect information about the quality of the data resulting from interviewer respondent interaction, for example, through an analysis of changed answers with regards to initial item non-response instances that later were completed.

\textit{Time measures for response latency measures.} Through CAI methods, it is possible to automatically record response latency measures. Timers are now embedded in audit trail files so their simultaneous analysis is possible. Additionally, timers –that can be measured to the millisecond– can be calculated with the level of detail as needed, ranging from section level times per item to the entire questionnaire. Traditionally, response latency measures have been used to identify problematic questions or cognitive issues (Bassili, 1996; Bassili & Scott, 1996), as well as to identify interviewers who may be speeding or not reading the full text of a question (Caspar & Couper, 1997; Olson & Peytchev, 2007).

An increasing number of researchers are interested in examining audit trails, and therefore some systems have started to produce more structured keystroke files; however they still present some challenges for their use as a tool to evaluate data quality.
Data Quality in Calendar and Time Diary Methods

In this dissertation I will focus on two types of nonsampling error: measurement and nonresponse, specifically item nonresponse. The first can also be categorized as an observational error; its sources can be the interviewer, the respondent, the questionnaire, or the method of data collection. The error of nonresponse, can be categorized—together with coverage and sampling errors—as a type of error of non-observation.

Measurement error. Before addressing each of these errors in the context of time diary methods, a short general discussion is in order. Additionally, it is important first to define measurement before discussing measurement error. Bohrnstedt (2010) defines measurement as “the assignment of numbers using rules that reflect or correspond to properties of a phenomenon or object”. And such “rules of correspondence between manifest observations and the numbers assigned to them define measurement in a given instance” (p. 349). Rules of correspondence, therefore, need to be as refined as possible, for the better the measures, the more accurate the assessment of the underlying relationship between variables will be (Bohrnstedt, 2010). Although the same definition can be used in the physical and the social sciences, the way measurement is applied is radically different. In the physical sciences, measures are based on standards that have been developed on theory and through experimentation. In the social sciences, such standards do not necessarily exist, or are very subjective at best. As a result, measurement error is a prevailing error transversal that impacts many disciplines and is of particular relevance to the survey science. That is why this type of nonsampling error seems to be the most studied across the survey literature (Groves, 1989; Biemer and Lyberg, 2003).
Measurement errors have also been categorized as observational errors, and thus can be understood as “deviations of the answers of respondents from their true values on the measure” (Groves, 1991, p. 1). If for whatever reason the respondent fails to provide the correct answer (e.g., failure to report amount of time devoted to child care the day before), then the answer provided deviates from the true value for that person. And if there is a tendency that this error happens across the entire population, the overall survey proportion will also be deviated from the population’s true value, resulting in measurement bias (Groves, 1991). The sources of observational or measurement error can include the respondent, the interviewer, the questionnaire, the mode of data collection, and the setting in which the survey is conducted. Respondent error occurs when the respondent intentionally or unintentionally offers inaccurate answers; depending on differences on cognitive abilities, characteristics (e.g., age, race) or motivations to respond, different respondents will provide information with different amounts of error. Interviewer errors result from the influence of interviewers on respondents’ answers while administering the survey. Interviewer error may occur when the interviewer fails to read the question correctly and completely, uses an intonation that may wrongly influence the respondent’s answer, probes inappropriately thereby biasing responses, or fails to record the answer correctly; interviewers can also produce error by deliberately falsifying data. Questionnaire or instrument error are the result of a poorly designed questionnaire that contains ambiguous questions, confusing instructions, and/or terms that are misunderstood by either the respondent or the interviewer. The mode of data collection can also produce measurement error; for instance, studies have shown that
information collected through the telephone can, in some cases, be less accurate and produce shorter answers than when the interview is conducted face to face (Groves, 1991; Biemer & Lyberg, 2003; Groves, 2004). Issues of social desirability also have an effect on measurement error: people tend to try to present themselves in the best possible light, so when conducting a survey on a sensitive topic (e.g., alcohol use), it has been shown that the setting or environment can have an influence on the accuracy on answers. For instance, it has been shown that when dealing with questions about drug abuse or sexual behavior, a more private setting can be conducive to more accurate results (Groves et al, 2009). Finally, Biemer and Lyberg (2003) also talk about measurement error that may arise from the “information systems” on which respondents base their responses. These can either be administrative records or the respondent’s own memory, and which may be erroneous, thereby producing answers that deviate from their true values.

As mentioned earlier, any error can either be identified as variance or bias. Measurement error in the form of variance will be generally less damaging for the statistic of interest than measurement bias. For instance, if all respondents’ answers to an income questions varies in a way that those who overreport their income balance against those who underreport their income, then errors are cancelled out, and the final estimate of the population mean income will still be unbiased (although individual true values will be biased). However, if those errors of those who underreport do not balance against those who overreport, then the final estimate of the mean income will be biased in a negative direction (Biemer & Lyberg, 2003, p. 43). At the same time, it has been pointed out that if, for instance, all persons underreport their income by say $10,000.00, though
the population mean income will be negatively biased, the correlations of reported income and other variables will not (Groves, 1991).

**Measurement error in surveys from a psychometrics perspective.** For a long time, the similarities and contrasts between measurement error in survey statistics and measurement error in psychometrics have been discussed. Importantly, it has been acknowledged that the theory of measurement from the fields of educational and psychological measurement are exceptionally suitable for the constructs used within the field of survey research, promoting a more in depth treatment of the concepts of reliability and validity (Bohrnstedt, 2010).

According to Groves (1991; 1989), the most important difference between statistical and psychometric approaches is that within survey statistics, the measurement problem lies in the operationalization of the question, that is, that the indicator itself can be weak; in psychometrics, on the other hand, the problem lies in the fact that the underlying, unobservable characteristic, can only be approximated through the application of a measurement. In other words, in psychometrics, the notion that the aim of the researcher is to measure an unobservable characteristic becomes more salient. More specifically, in the psychometric literature, it is acknowledged that there can be a difference between the observed response (i.e., the number assigned to the observable characteristic) and the underlying unobserved variable that generated the response. This difference between observable and unobservable variables is defined as measurement error. Such differences can be ‘internal’ (e.g. lack of motivation to respond) or external (e.g., an interruption while responding) (Bohrnstedt, 2010). That is why in Classical True
Score theory, for a variable $x$ that is measured across persons, the relationship between the observed and the true scores is:

$$x = \tau + \varepsilon$$

Where $x =$ observed score; $\tau =$ unobserved true score; and $\varepsilon =$ unobserved measurement error.

Psychologists and survey methodologists have found the need to distinguish between verifiable and unverifiable true scores. Verifiable (at least in theory) or Platonic true scores correspond to behavioral responses (e.g., number of drinks per day, whether a person voted in the last election or not, etc.), while phenomena such as psychological states, attitudes, values, and beliefs belong to classical true scores. In the classical true score theory model (CTST), the true score is defined as the expected value of the observed score, from which the mean (or expected value) of the errors of measurement across the entire population is zero. This true score model has become a common measurement error model in the survey literature.

**Classical true score model (Model 0).** The true score model (also referred to as Model 0) posits that the observation of a characteristic of a randomly selected individual equals to the sum of two terms: a “true value” and an error term (Biemer & Stokes, 1991). The mathematic notation of Model 0 is:

$$y_j = \mu_j + \varepsilon_j$$

where $y_j$ is a single observation of a particular characteristic of individual $j$, composed of the true value $\mu_j$ and an error term $\varepsilon_j$.

The assumptions of model 0 are the following (Biemer & Stokes, 1991):
a. The sample has been selected using simple random sampling without replacement (SRSWR). The implications of having SRSWR include that the expected value of $\mu_j$ is the sum of the characteristic of interest across all members of the population divided by the population size \( E(\mu_j) = \mu = \frac{1}{N} \sum_{i=1}^{N} \mu_i \). Likewise, the variance of $j$’s true value \( Var(\mu_i) \) equals to the variance of the true value among the population \( \sigma^2_{\mu} = \frac{1}{N} \sum_{i=1}^{N} (\mu_i - \mu)^2 \).

The next set of assumptions relate to the fact that within each individual respondent $j$ there is an error distribution of all the possible responses. Thus the error associated to the response individual $j$ gives to a question in a particular trial of the survey (a survey that can be conceptually repeated over and over again), is just one of the many random errors than could have been selected from an infinite number of errors from individual $j$.

Therefore, the observed value $y_j$ is the result of the true value of individual $j$ ($\mu_j$) plus an error term that makes the observed value to deviate from the true value. Therefore:

b. The expected error, at the individual level, -that is, conditional on each respondent $j$- is zero \( E(\varepsilon_j|j) = 0 \). This means that the response deviations are not associated with the true value.

c. The variance of the error term given individual $j$, equals to the variance of the distribution of responses of individual $j$ over all possible (conceptual) repetitions of the survey. In other words, it is assumed that there exists variability within respondents. \( Var(\varepsilon_j|j) = \sigma_j^2 \)
d. The covariance between the response deviations of respondent $j$ and respondent $j'$ is zero. That is, the response deviation of one individual in the sample is not related to the response deviation of another individual in the sample. \( \text{Cov}(\varepsilon_j, \varepsilon_{j'}) = 0 \text{ for } j \neq j' \)

e. If we were to reinterview respondent $j$, it is assumed that response deviations given in the initial interview ($\alpha$) and the reinterview ($\alpha'$) are not correlated. 
\( E(\varepsilon_j, \varepsilon_{j'}|j) = 0, \alpha \neq \alpha' \)

f. Measurement error adds up

g. A true value of the characteristic of interest exists for each individual $j$. 

Assuming Model 0 to be true, Biemer and Trewin (1997), discuss the effect of measurement error on univariate statistics (i.e., means, totals, proportions). Given that Model 0 is a model of uncorrelated error effects, which can be expressed as follows:

\[
y_{ij} = \mu_{ij} + d_{ij} \\
= \mu_{ij} + b_i + \varepsilon_{ij}
\]

In this case, we are looking at potential correlations between the errors due to interviewers ($i$) and other sources of error such as the respondent ($j$). The expression above is the model for the $ij^{th}$ observation, where $ij$ denotes the observation of the $j^{th}$ unit in a set of units assigned to the $i^{th}$ interviewer. The deviation $d$ is composed of two error terms, $b_i$ and $\varepsilon_{ij}$. In this case, $b_i$ denotes the interviewer effect, which is the same across all units of the $i^{th}$ interviewer's assignment. It has mean $B_b$ and variance $\sigma_b^2$. $\varepsilon_{ij}$ represents the errors arising from the respondent or any other source of error. These errors are
assumed to be random variables, with a mean $B_e$ and variance $\sigma^2_e$. $\mu_{ij}$ is the true value of the $j^{th}$ unit within the assignment of the $i^{th}$ interviewer.

In Model 0, it is assumed that there are no interviewer effects. That is, there is no variance in the responses due to interviewers. As a result, variance $\sigma^2_b$ is zero the covariance of the deviations of respondent $j$ and respondent $j^\prime$ is zero.

Under the previous assumptions and SRSWR, the estimator for a univariate statistic such as proportion $\bar{X}$ is (Biemer & Trewin, 1997):

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{m} y_{ij}$$

The expected value for $\bar{y}$ is:

$$E(\bar{y}) = \bar{X} + B_d$$

where $B_d$ is the bias in the sample mean, and $B_d = B_b + B_e$. In model 0, the error distributions of $B_b$ and $B_e$ are equal to zero, and so $B_d = 0$. As a result: $E(\bar{y}) = \bar{X}$. In other words, even in the presence of measurement error, when errors are uncorrelated and randomly distributed (as in Model 0), the sample mean is an unbiased estimate of the population mean (or any other univariate statistic of interest).

However, as Biemer and Trewin (1997) explain, the effects when a regression model is involved may be different assuming Model 0 to be true. For instance, if we have the following regression equation:

$$y_{ij} = \beta_0 + \beta \mu_{ij} + \xi_{ij}$$

where $\beta_0$ is the intercept, $\beta \mu_{ij}$ is the slope of the regression model, and $\xi$ are independent, identically distributed errors of the model, with a mean of 0 and a variance.
of $\sigma^2_\xi$. Under model 0, the measurement error $d_{ij}$ is assumed to be normally distributed and independent of the model errors $\xi$.

The expected value of the slope coefficient $\hat{\beta}$ is a function of the reliability ratio $E(\hat{\beta}) = R\beta$. The reliability ratio $R$ under model 0 is the ratio of the variance of the true values over the variance of a single observation. Since under measurement error the reliability ratio will always be less than 1, under Model 0, the estimator of the slope will be biased toward zero, or what is also known as being “attenuated” (Biemer & Trewin, 1997). The line with (uncorrelated) measurement error will be less steep than the true regression line. As for the intercept $\beta_0$, since it is not what would the sign of the bias of the sample mean will be, with uncorrelated measurement error the intercept will be biased, but it is not possible to know in which direction.

Platonic scores model (Model for acknowledging the possibility of bias).

Classical true score theory becomes insufficient for most survey applications because of the need to acknowledge possible biases in survey measurements (Groves, 1991), and this cannot be accomplished if non-Platonic true scores are assumed. Thus, one can only speak of bias by assuming Platonic true scores, where

$$x = \tau^* + \varepsilon$$

In this case, the means of the errors of measurement may or may not be zero. That is, in any given situation, the measurement of $\tau^*_i$ may have error and the expected value of the observed score will not equal the true score, and the expected mean of the errors of measurement may be different from zero. Response bias $\beta$ can arise, and is defined as (Bohrnstedt, 2010):
\[ \beta = E(\varepsilon) = E(\tau^*) - E(x). \]

Additionally, with the assumption of Platonic true scores it is possible to test the independence of true scores from their errors by computing the covariance of \( \tau^* \) and \( \varepsilon \) across the population. If this correlation is found to be nonzero, then the true scores and errors are not independent. In this case, the variance of \( x \) is computed as follows:

\[ \sigma_x^2 = \sigma_{\tau^*}^2 + \sigma_\varepsilon^2 + 2C(\tau^*, \varepsilon). \]

Where \( \sigma_x^2 \) is the variance of the observed responses; \( \sigma_{\tau^*}^2 \) is the variance of the true scores; \( \sigma_\varepsilon^2 \) is the variance of the errors; and \( 2C(\tau^*, \varepsilon) \) is twice the covariance between the true score and the corresponding measurement error (Bohrnstedt, 2010, p. 351). Besides allowing for the possibility of bias, the assumption of Platonic scores allows the testing of many definitions and assumptions in classical test score theory. The limitation stems from the fact that is rarely the case that one has trustworthy verifiable measures, even when dealing with for behavioral variables, especially as a result of social desirable responses (e.g., underreport alcohol use; overreport voting behaviors).

In this case, where bias can exist, the mean of the errors of measurement will be nonzero, and thus the effect of error on univariate statistics (i.e., means, proportions) will be that the statistic of interest will be biased. However, bivariate relationships (i.e., correlations, regressions) can remain unaffected, as long as it can be assumed that the errors of measurement are uncorrelated with the true scores. However, if the mean of the measurement errors is not equal to zero as a result of social desirability, the true scores and the errors of measurement may not be independent and, as a result, all the statistics can be biased (Bohrnstedt, 2010, p. 351).
Reliability and validity. Reliability and validity are better understood in the context of Model 0, where only variable errors exist (Groves, 1991). Reliability has been defined as the ratio of the true score variance to the observed variance, that is, “the extent to which the variance of an observed $x$ is due to systematic sources rather than ‘noise’” (Bohrnstedt, 2010, p. 352).

Therefore:

$$\rho_x = \sigma_t^2 / \sigma_x^2.$$ 

Because in this case variance refers to variability over individuals in the population and over trials within a person, the concept is not defined for measurements on a single individual, but on an entire population (Groves, 1991).

As for the concept of validity, many definitions of validity have been put forward, but a general definition is that “validity indicates the degree to which an instrument measures the construct under investigation” (Bohrnstedt, 2010, p. 373-374). Therefore, the correlation between the true score and the observed score is the theoretical validity of $x$, as it measures the extent to which an observed item correlates with the (latent) construct of interest (Bohrnstedt, 2010), or the correlation between the true score and the respondent’s answer over trials (Groves, 1991). Given that validity is based on correlations, in this case also, it can only be defined on a population and not on a single person (Groves, 1991). Additionally, since in true score theory it is assumed that errors are uncorrelated with the true values of the respondents on any of the trials, it follows that the theoretical validity of a measure is nothing more than the square root of its reliability:

$$Validity = \sqrt{\rho_x}.$$
Because of this, theoretical validity cannot be equated to unbiasedness as understood in the survey literature. Indeed, it is well known that no measure can be valid without also being reliable, but a reliable measure is not necessarily a valid one. However, when it comes to relate unbiasedness with variance and bias, things are not so straightforward. As Groves (1991) puts it, “a sample statistic may have an expected value over samples equal to the population parameter (unbiasedness), but have very variance from a small sample size. Conversely, a sample statistic can have very low sampling variance (from an efficient sample design) but have an expected value very different from the population parameter (high bias)” (p. 9).

The differentiation between theoretical and empirical validity is very important. Whereas theoretical validity is the correlation between the underlying latent construct and the observed measure, empirical validity is the correlation between the observed measure and another observed criterion than purportedly measures the same construct. In the area of psychological and educational measurement, the traditional view has been to divide validity into four different types: content validity, criterion validity, construct validity, and convergent and discriminant validity (Bollen, 1989). According to Messick (1995), such a view is “fragmented and incomplete”, as it fails to “take into account both evidence of the value implications of score meaning as a basis for action and the social consequences of score use” (p. 741). Messick proposes a new unified concept of validity, where those previously separated interpretations are interrelated as elements of a comprehensive theory of construct validity. In this new unified theory of validity, elements of content, criteria, and consequences are included. Messick thus defines
validity as “an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment” [emphasis added] (p. 741). This definition implies that what is validated are the interpretations of the scores or other assessment tools, and the uses of those scores for particular applied purposes (Thorndike & Thorndike-Christ, 2009). Importantly, and drawing a parallel with the concept of data quality understood as fitness for use, Thorndike and Thorndike-Christ (2009) argue that there are two type of inferences that need to be validated: interpretive inferences and action inferences. The first ones refer to what test scores mean; the latter refer to the appropriateness and utility of test scores as the basis for some specific action. Importantly, this validity theory goes beyond scores and their meaning, and includes “the value implications of score interpretation and the utility, relevance, and social consequences associated with test use” (Thorndike & Thorndike-Christ, 2009, p. 179).

Thus, for Messick (1995), what needs to be valid is the interpretation of the scores and any action implications, and not the scores themselves. Importantly, Messick argues that this notion of validity apply to any type of assessments (questionnaire, observations, etc.), and not only test scores. In sum, Messick understands validity in a broad sense, in which construct validity (i.e., whether a measurement corresponds to some meaningful trait or construct that we are trying to measure) subsumes other types or aspects of validity.

In Messick’s theory, the main threats to (construct) validity include construct underrepresentation and construct-irrelevant test variance. The first threat refers to the fact that it can be the case that the measurement tool (test, questionnaire, etc.) is too
narrow and does not include central elements of the construct. The second threat refers to the presence of reliable variance that does not belong to the construct being measured (Thorndike & Thorndike-Christ, 2009, p. 182).

Assessing measurement error in surveys using reliability and validity measures.

The empirical literature shows that measurement error assessment in surveys has included reliability and validity assessments, sometimes jointly and sometimes separately. One of the most common ways to assess measurement error is one where repeated measurements of the same persons take place (Groves, 1991). In this case, a person is re-interviewed and a question is asked a second time. Or, two equivalent measures are utilized in the same test or survey, or in a different test or survey implemented at a later point in time. In the educational measurement parlance, this is simply a measure of consistency or, as it is more popularly known, a test-retest measurement study. This type of assessment looks at how stable is the measure, but not necessarily as how valid it is. In this case, it is assumed that the items will correlate between each other because both measure the same underlying unobserved true variable $\tau$.

The equations used in this case are:

$$x_1 = \tau + e_1; \ x_2 = \tau + e_2$$

It is assumed that the errors between $x_1$ and $x_2$ are uncorrelated, that their expected value across respondents is zero ($E[e_{i1}] = E[e_{i2}] = 0$. It is also assumed that both variables have equal variances ($\sigma^2_{e_1} = \sigma^2_{e_2}$), and therefore that errors are uncorrelated over trials ($\text{Cov}(e_{i1}, e_{i2}) = 0$). In the survey context, however, such assumptions are difficult to
maintain mainly due to memory effects. For instance, memory from the first trial answer may prompt the respondent to give exactly the same answer (to look consistent), or to give a different answer (to show flexibility) (Groves, 1991, p. 18). In that case, the covariance between errors of the two items are no longer uncorrelated. Likewise, on the interviewer side, knowledge of the previous answer and introduce correlation of errors across trials.

The analysis of validity is much more complex, as according to many authors (e.g., Alwin, 2010) it involves a philosophical debate. For instance, Alwin (2010) argues that validity has to do with the relationship between a latent variable T (true value) and a theoretical construct of interest; in that sense, the estimation of validity is a very elusive endeavor, and all we can aim at is at seeking some type of evidence of validity. In any case, other authors, beginning with Campbell & Fiske (1959) have argued that assessments of both reliability and validity are possible. One way of doing so is through the multitrait multimethod (MTMM) design. In such design, construct validity is detected when two (or more) different measurements of the same trait observed through maximally different methods show correspondence; reliability is demonstrated when there is correspondence between two measurements that assess the same trait using maximally similar measures (D. T. Campbell & Fiske, 1959). The implicit measurement model in a MTMM design has also been called the common factor model (Groves, 1991), because multiple indicators, each one with a specific error (method/interviewer effect).

5 By the same token, according to Alwin (2010) reliability refers to the relationship between an observed measure X and a latent variable T being measured. Accordingly, reliability is still measured through the correlation between two identical measures (x₁ and x₂) of T (the true value), just as suggested above.
and random error, are used to measure an underlying factor or trait. In this model, it is possible to empirically estimate some of the components of measurement error, such as the amount of variance due to the true scores, to method effects, to interviewer effects, and to random error.

Another design used to study validity is the so-called record check or validation study. In it, survey reports are compared to a different external dataset (e.g., administrative records) that may contain the same information asked in the survey in question, but that is believed to have more accurate information. The main limitations of these type of studies include the fact that the information contained in the records is not necessarily without error, and that only a limited number of variables can be found in administrative record systems (Groves, 1989).

A third type of validation study is the “gold standard” designs. In these, responses obtained through a particular “new” method (e.g., event history calendar), are compared to responses obtained through a different method approach that is considered the accepted or gold standard (Belli, Shay, & Stafford, 2001; Bilgen & Belli, 2010). The validity of the “new” method will be gauged vis-à-vis the gold standard approach (Alwin, 2010). The limitation here again is that to begin with, one cannot be sure that the gold standard is actually validly measuring the underlying construct, and one may argue that what is being really being assessed is reliability.

Beyond classical test theory: “New” methods to establish validity and reliability.

Because of the difficulty to establish validity, researchers evaluating measurement error have commonly employed internal consistency approaches based on the CTST to assess
the reliability of the data. A very popular approach within the CTST has been use of the Guttman-Chronbach Alpha coefficient. Coefficient Alpha has been used to assess reliability in cross-sectional survey data by estimating the reliability of a linear composite score made up of multiple measures of a given concept (Alwin, 2010). Several shortcomings have been noted with the use of this coefficient. First, it assumes Tau-equivalency, which means that all items or questions equally relate to the unobserved factor (i.e., “true-score equivalency). Second, Coefficient Alpha assumes unidimensionality, that is, all measures should reflect one single underlying variable (which may not be the case in many situations). Third, it assumes that the errors in the measures are independent of one another (Alwin, 2010). Finally, the main problem of Coefficient Alpha, and the internal consistency approaches in general, is that they are generally used to assess the reliability of linear composites, which is not a useful way to evaluate individual survey questions (Alwin, 2010).

CTST can be viewed as a one model (with a set of restrictive assumptions) within the more general family of Latent Trait Measurement Models (LTMM). Other models that make part of the LTMM family include, for instance, Item Response Theory (ITR) Models, and Confirmatory Factor Analysis Models (CFA), of which the MTMM is an example (Maruyama, 1998). In any case, given the limitations of CTST, some efforts have been put forward by (few) researchers to evaluate measurement error using other more general techniques that allow the measurement of latent traits, relaxing some of the assumptions of CTSC. For instance, Andrews (1984) estimated the amount of variance due to the true value, to random error, and correlated errors (e.g., method effects) in order
to establish the “measurement quality” of survey items through a Confirmatory Factory Analysis model\(^6\). Likewise, (Saris & Andrews, 1991) evaluated survey questions through a MTMM design (i.e., a type of CFA)\(^7\), where they estimated validity coefficients, method effects coefficients, and reliability coefficients.

Another important effort beyond the use of CTST to estimate measurement error is one by Biemer (2011) where he uses Latent Class Analysis (LCA) to fundamentally model and estimate classification errors in surveys. He also uses LCA to estimate reliability in survey items.

**Nonresponse error in surveys.** Nonresponse error comprises unit nonresponse, item nonresponse, and incomplete responses (Biemer & Lyberg, 2003). Unit nonresponse exist when a sampling unit (e.g., household individual within household, business) fails to respond to the entire questionnaire. Item nonresponse exists when the respondent skips or refuses to respond some items, and only partially completes the questionnaire. A typical example of item nonresponse is the question on household income which tends to be left blank even though other sections of the interview are readily completed. A final type of nonresponse error are incomplete responses to open-ended questions; this occurs when the respondent provides a response, but it is too short or insufficient, so that it cannot be classified. The typical example in this case are the labor force open-ended questions about occupation, where the respondent may provide some but insufficient

---

\(^6\) Though he calls it a “Structural Equation Model”. I argue this is not the case, because he did not try to establish relations between the latent variables that were measured.

\(^7\) These authors also contend that they are using a structural modeling approach, but again, no relations between latent variables were established.
information about his occupation, that prevents the interviewer or coder to assign a correct industry or occupation code.

Nonresponse can lead to variance or bias, though the risk of variable error is low while the risk of systematic error (bias) is high. However, it has been shown that even though a survey may a low response rate, nonresponse bias will only occur if those who do not respond are different from those who respond in the variable of interest (Groves, 2006).

**Influences on the decision to respond to a survey question.** When respondents are asked a survey question, they have first two questions to answer to themselves: whether they can respond or not, and whether they will respond or not. If they decide negatively in either case, item nonresponse will be the result (Beatty & Herrmann, 2002). Item nonresponse creates analytical difficulties: on the one hand, statistical analyses are compromised, as item nonresponse reduces the effective sample size; on the other hand, imprecise or incomplete answers create a dilemma for the researcher in that it is difficult to gauge the extent to which the quality of the response has been compromised.

Theory on item nonresponse regarding autobiographical questions indicate that there are two fundamental broad influences on the decision to respond: (a) cognition, i.e., the ability to remember and use the relevant information to the task at hand; and (b) motivation, which can in turn be affected by sensitivity issues (if the respondent feels the question may compromise how others look at him), burdensome questions (if the items ask for information that is hard to provide), or conflict of interest issues (if the respondent feels the information can be used against him) (Beatty & Herrmann, 2002).
Additionally, features of the questions and questionnaires themselves can influence the decision to respond. For instance, how clear is the question or how complex is the response task can determine the extent of the cognitive challenge facing the respondent, and thus affect his motivation to respond. Beatty and Herrmann (2002) propose that the decision to respond or not is fundamentally driven by the following factors: (a) cognitive state (the extent to which the information is available in the respondent’s memory: available, accessible, generatable, inestimable); (b) adequacy judgments (the respondent’s perception of how accurate the answer needs to be); and (c) communicative intent (the respondent’s motivation to actually report the answer) (p. 72-73).

**Nonresponse error in time diary methods.** The quality and validity of reports also hinge upon the way the information is collected. When studying the specific ways in which people use their time, conventional standardized questionnaires have been used. For instance, respondents may be asked to estimate how much time they allocate to their different activities using a stylized list of activities. Yet, it has been found that it can be very difficult for respondents to produce accurate responses using this approach and measurement error is likely to arise. From the time use research perspective, two different mechanisms can be associated with error when using stylized questions: (a) the lack of flexibility that prevents the conversation to flow in a natural way (Jabine, Straf, Tanur, & Tourangeau, 1984; Suchman & Jordan, 1990), and (b) the fact that activity frequency and duration surveys using stylized sets of possible activities provide reports that are episodic and may be taken out of context (Pentland et al., 1999). Time diaries and calendar-based
interviews, which ask about time-use using a conversational approach, have been proposed as an alternative to overcome such complications, mainly because of their ability to encourage respondents to incorporate in their cognitive processing temporal changes that serve as cues that assist providing a more accurate reporting of events (Belli et al., 2009a).

The reduction of error in timeline surveys

The role of flexible conversation techniques in reducing error in timeline surveys. Conventional standardized interviewing is the most widely practiced technique as it purportedly reduces variance in responses due to interviewers and maximizes variance attributable to the actual differences in respondents’ circumstances (Fowler & Mangione, 1990; Fowler & Cannell, 1996). The first mechanism that may produce error in retrospective survey reports, namely the lack of conversational flexibility can be attributed precisely to the standardization of the questionnaire. Time diary and calendar methods address this source of error by allowing interviewers to lead the conversation in a natural manner (Houtkoop-Steenstra, 2000; Maynard & Schaeffer, 2002). Although in conventional interviewing the wording of questions is standardized there is no guarantee of a non-ambiguous and consistent understanding of questions by respondents (Houtkoop-Steenstra, 2000; Suchman & Jordan, 1990). By assuming an interview is nothing more than a neutral measurement instrument, conventional standardized interviewing suppresses the elements of ordinary conversation, compromising both the understanding of the intended meaning and the validity of the answers.
Numerous studies have demonstrated how interviewers frequently cannot maintain the rules of standardized interviewing (Belli, Lepkowski, & Kabeto, 2001; Belli, Bilgen, & Al Baghal, 2013; Houtkoop-Steenstra, 2000). The conversational interviewing technique accepts that an interview involves an interaction between the participants, in which the rules of conversation will in some manner be present (Schwarz, 1996; Schwarz, 2009). Following Clark and Schober (1992), this technique recognizes that language is not about the literal meaning of words but about people and what they mean. The coveted goal of providing greater consistency to question meaning may be better reached by allowing interviewers to clarify the concepts and assist respondents when doubts of any sort arise (Conrad & Schober, 2000; Schober & Conrad, 1997).

Calendar and time diary methods take advantage of the conversational survey technique by disregarding the need to use fixed words and phrases and permitting flexibility to interviewers as long as they complete the diaries or the calendars in the way they are intended. The benefits to data quality due to the use of the conversational technique in diaries and calendars is further enhanced by the memory cues that are encouraged.

_The role of (autobiographical) memory in reducing error in timeline surveys._

Answering survey questions necessarily involves cognitive and memory processes and their limitations are associated with error in survey reports (Belli, 2013). For that reason, survey methodologists have incorporated cognitive science perspectives into their field of

---

8 This section draws heavily on Belli, 2013.
study (Jabine et al., 1984; Tanur, 1992)\textsuperscript{9}. An example is the classic question response model proposed by Tourangeau (1984), which involves question comprehension, memory retrieval, the judgment of the relevance of retrieved information, and the selection and editing of the final response (Tourangeau, Rips, & Rasinski, 2000). Although the role of memory is important in the answering of any question, in the case of time studies as time diaries or calendars, the role of memory is crucial as respondents are queried about past events that they have experienced. For instance, the American Time Use Survey (ATUS), which is a time diary survey conducted by the US Census Bureau to understand the time use patterns of the population of the United States, asks the following question: “Now I'd like to find out how you spent your time yesterday, from 4:00 in the morning until 4:00 AM this morning. I'll need to know where you were and who else was with you.” Likewise, a great reliance on memory processes will occur with questions asked about longer periods of time, such as with the Panel Study of Income Dynamics, a prospective national study of life course socioeconomics and health, which asks the following question: “I'd like to know about all of the work for money that you have done since January 1, [Past year]. Please include self-employment and any other kind of work that you have done for pay.” In both circumstances, respondents will need to retrieve information from their autobiographical memory. Importantly, if the past 24 hours or the past year consist of complex experiences by including several different activities or jobs,

\textsuperscript{9} The initiative to incorporate cognitive science knowledge into survey research started in the 1980’s was called the Cognitive Aspects of Survey Movement (CASM) and continues to this day. For a brief history on the topic see the Preface of Tanur (1992).
respondents will have to make a considerable memory effort to derive a complete and accurate answer.

Belli and colleagues have noted that calendar questionnaires encourage the use of cues that exist in the structure of autobiographical knowledge which, together with flexible interviewing, enhance the quality of retrospective reports in comparison to conventional standardized questionnaires (Belli, 1998; Belli, Shay, & Stafford, 2001). Importantly, they have shown that improvements in retrospective reporting also occur when collecting subjective assessment information, such as health status over the life course (Belli, Agrawal, & Bilgen, 2012). In particular, they have noted that events that are more easily remembered can become memory cues that will help respondents to remember events that are more difficult to remember (e.g., see Belli, 2013). There are three possible types of cueing: top-down, sequential, and parallel cueing. Top-down cueing occurs when more general events serve as cues to remember more specific ones: remembering the name of an employer helps remembering more specific details such as weekly pay. Sequential cueing occurs when a remembered event is used as an anchor that aids in the remembering of a temporally adjacent event within the same life domain: remembering that one worked for one employer during a period of time helps remembering the name of the employer one worked with afterwards. Finally, with parallel cueing, a remembered event in one life domain assists in the remembering of an event from a different life domain that occurred contemporaneously or nearly so: remembering a change of residence helps remembering that one changed employment.
Time diaries also take advantage of these cueing properties, although differently. Whereas calendars encourage respondents to report about periods of stability and points of transitions in different life domains (work, relationships, health), time diaries ask about transitions between the different activities one engaged in during the day. In calendar surveys, respondents provide information about a number of timelines covering events from the different life domains of interest, and reference periods can range from months, to years, and up to the entire lifetime. In time diaries, respondents provide information about activity sequences, and the context in which these occurred; their reference period is generally 24 hours. In terms of cuing, given that time diaries are driven by location, transitions between activities generally involve a change in location (within the house or traveling to a different place; see Stafford, 2009). Hence, top-down cuing may occur in which the more general event (the activity) will trigger one to remember the more specific detail of ‘where’. A bottom-up cueing can also occur as one may first remember ‘where’ before remembering the activity. Likewise, sequential cueing may occur regarding the details and context of the activity in which remembering one activity may assist in the remembering the next. Finally, parallel cueing may occur when a person reports a secondary contemporaneous activity.

Research has found that timeline methods (especially calendars) do enable more complete reconstructions of one’s past and an enhancement in retrospective reporting data quality (Belli, Lee, Stafford, & Chou, 2004; Belli et al., 2013; Bilgen & Belli, 2010). Given that social, behavioral, and health scientists will continue to administer surveys that ask respondents about their pasts, such interviewing methods are encouraged in order
to produce more valid scientific inferences about individual life course trajectories and social interrelationships.

**Satisficing and the Creation of Rapport (in Time Diaries)**

As established by previous research, a key determinant of data quality in surveys is question difficulty, which significantly interacts with the cognitive ability of the respondent (Knauper, Belli, Hill, & Herzog, 1997). In order to provide an optimal answer, the respondent needs to perform the cognitive tasks of comprehending the question, retrieving the information in one’s memory, judging the relevance of the retrieved information, and selecting and editing the final response (Tourangeau et al., 2000). The difficulty of these cognitive tasks will depend on the content or topic of the survey question, the time frame it is referred to, or the response options that are provided (Knauper et al., 1997). By extension, in time diaries, where no scripted questions are used, it has been shown that more complicated interviewing situations will lead to poorer data quality than less complicated ones (Belli et al., 2013).

As pointed out by Krosnick (1991), difficult questions may lead respondents to “satisfice”, that is, to provide an answer that externally satisfies the survey task at hand, but that not necessarily means the respondent provided the “optimal” response. In Krosnick’s theory, satisficing occurs when respondents fail to devote the necessary amount of energy and effort in order to efficiently perform each of the steps of the survey response process. When, on the contrary, respondents do so in a careful and thorough manner, “optimizing” occurs. For Krosnick, optimizing and satisficing are part of a continuum indicating “the degrees of thoroughness and bias in retrieval and integration”
(Krosnick, 1991, p. 215). Thus, satisficing can be attributed to incomplete or biased retrieval and judgment, or the elimination of any of these cognitive stages altogether. Additionally, Krosnick suggests that the likelihood that a respondent will satisfice when answering a question is a function of the following factors (and their interactions): (a) Inherent difficulty of the task, (b) Respondent’s ability, and (c) Respondent’s motivation. Given these notions, it would be expected that respondents with low cognitive ability and/or motivation will be more likely than those with higher ability and/or to provide incomplete or biased responses.

It has long been recognized that the attributes of either the respondent or the survey design (e.g., questionnaire, interviewers, setting, etc.) can influence each of the stages of the response process. That is, any of the survey attributes may affect not only the motivation and the level of effort interviewers and respondents are willing to undergo to complete their tasks as accurately as possible, but they may also affect interviewers’ and respondents’ perceived difficulty (i.e., the cognitive or psychological burden) of the survey task (Tourangeau et al., 2000). On his part, Krosnick (1999) has posited that respondents will frequently try to minimize those costs through strategies by which they can shortcut the cognitive process required to produce accurate reports.

In the context of time use surveys (such as calendars or time diaries), it has been suggested that the response process may benefit from increased respondent motivation because of the conversational interaction that develops (Belli, R. and Callegaro, M., 2009). In fact, optimizing strategies might actually occur, as “calendar interviewing encourages coherence and precision in reporting what happened and when, and because
conversational flexibility and effective retrieval cues enhance the motivation, engagement, and interest of the respondents” (p.37).

Further, such conversational interaction may foster the creation of rapport (i.e., the attempt to create a positive and friendly atmosphere during an interview between the interviewer and the respondent to enable a more productive interaction). It is important to mention that especially in the context of standardized interviewing, the concept of rapport is a much contested one, as it is argued that this rapport, that results in the creation of a closer relation between the respondent and the interviewer, may actually bias responses as respondents may feel more pressure to put themselves always in the best light possible. However, in the context of conversational approaches such as diaries or history calendars, things may be somewhat different. For instance, Belli et al. (2013) have shown that when it comes to retrospective reports, rapport is usually helpful, but that it may be detrimental if the information being queried is sensitive in nature. In any case, it is accepted that the mere presence of an interviewer has an effect on respondents (Krysan and Couper, 2003).

In other words, the involvement of an interviewer is bound to produce complications in the interaction with the respondent (Houtkoop-Steenstra, 2000; Suchman and Jordan, 1990). As it has been argued by some, conversational flexibility will facilitate the interaction and help repair any type of miscommunication. At the same time, however, it has been established that flexibility also has disadvantages compared such as the increased difficulty in evaluating interviewer behavior and interviewer training (Houtkoop-Steenstra, 2000), and that conversational interviews take much longer to complete (Schober and Conrad, 1997).
The Use of Paradata to Assess Data Quality and Identify Interview Rapport

It has been argued that paradata, data about the data collection process, may be useful to reflect the cognitive tasks involved in the survey response processes. In fact, paradata not only have the “potential to shed light on the survey process itself”, but “with proper ‘mining’ they can point to errors and breakdowns in the process of data collection” (Kreuter, 2013, p. xv). In that sense, the use of paradata, when examining survey data quality from a TSE perspective, is based on the notion that error can occur when there is a breakdown in the cognitive response process (Olson and Parkhurst, 2012), and that paradata is deemed capable of capturing such breakdowns (Olson & Parkhurst, 2013). Under this same argument, one may argue that paradata can help uncover interview rapport, as paradata are also capable of reflecting the way the interview was conducted, for instance by offering information on response times, edits to answers, and call-backs.

The types of paradata that that have been used to examine survey data quality include response times (measured in milliseconds and captured automatically by the CATI or CAPI instrument), mouse clicks (which may reflect back-ups and answer changes), keystrokes or audit trails (which reflect the use of specific keys), and call and case history files (which include, for instance, the number of attempts before obtaining a complete interview and the number of completed interviews). Response times have been the most widely paradata variable used so far, particularly for the study of measurement error. For instance, some studies have shown that longer response times may be a sign of
difficulties in the response process, or of potential problems with survey questions
(Bassili & Scott, 1996; Yan & Tourangeau, 2008).
CHAPTER 4
THE AMERICAN TIME USE SURVEYS (ATUS)

The ATUS is a federally funded nationally representative time-use survey, sponsored by the US Bureau of Labor Statistics, in which the data are collected by the US Census Bureau. The ATUS started in 2003, and is the first ever survey of this type undertaken in the United States. It is an ongoing survey that provides annual estimates of the how people in the United States spend their time. The most common activities reported include paid work, child care, volunteering, and socializing; these data, including microdata files, are available for public use on the BLS Website (Phips and Vernon, 2009, p. 109). In the ATUS, approximately 24,000 individuals are sampled each year from the pool of the final wave of respondents from the Current Population Survey (CPS). For this dissertation, I will use data that come from the 2010 ATUS, which has a response rate is around 56.9% (AAPOR RR 2).

The ATUS data is collected throughout the entire calendar year in order to control for bias that may occur because of seasonal effects. Likewise, weekends are oversampled in order to allow for similar precision of estimates from workdays (ATUS User’s Guide 2011). Households with telephone numbers are sent an advance letter and then contacted and interviewed by telephone. Households without telephone numbers are sent a letter with a call-in toll free number and a $40 incentive is sent to them. Any household member 15 years or older is eligible for selection. The interview can be conducted in either English or a different language. The ATUS interview includes several sections besides the time diary portion, including: the household roster; employment history;
summary questions (work and income-generating activities, secondary childcare, volunteering); trips history; labor force status; and earnings and school enrollment.

When interviewed, respondents are requested to report all activities they engaged in the day before starting at 4 AM of the previous day until 4 AM of the interview day. The respondent is to report not only all activities, but also their timing and additional details such as the place of occurrence and whether somebody else was present or not. Interviewers use predefined codes to record the reported activity, the place where the activity occurred, and with whom. After data collection, activities reported by respondents are coded using a three-tier coding system. The first two digits represent the major activity category (18 codes); the next two digits represent the second level of detail, and the final two digits represent the third, most detailed level of activity. The ATUS also codes six types of errors at the activity level: missing travel or destination, insufficient detail in verbatim, recorded simultaneous activities incorrectly, refusals to report an activity, recall failure, inability to code activity at 1st tier.

The ATUS Wellbeing Module

In years 2010, 2012, and 2013, the ATUS Wellbeing Module was fielded by the US Census Bureau. The Wellbeing Module asked respondents about how they felt during selected activities, as well as general health information. To be selected for this module, the activity had to be at least 5 minutes in duration and could not be any of the following activities or have any of the following codes: Sleeping (0101xx); Grooming (0102xx); Personal Activities (0104xx); Don’t know/Can’t remember (500106); and Refusal/None
of your business (500105). Questions 1 through 7 of the wellbeing module were posed as follows (U.S. Bureau of Labor Statistics, 2014):

Now I want to go back and ask you some questions about how you felt yesterday. We’re asking these questions to better understand people’s health and wellbeing during their daily lives. As before, whatever you tell us will be kept confidential. The computer has selected 3 time intervals that I will ask about.

Between [STARTTIME OF EPISODE] and [STOPTIME OF EPISODE] yesterday, you said you were doing [ACTIVITY]. The next set of questions asks how you felt during this particular time.

Please use a scale from 0 to 6, where a 0 means you did not experience this feeling at all and a 6 means the feeling was very strong. You may choose any number 0,1,2,3,4,5 or 6 to reflect how strongly you experienced this feeling during this time.

1. Happy  First, from 0 – 6, where a 0 means you were not happy at all and a 6 means you were very happy, how happy did you feel during this time?
2. Tired  From 0 – 6, where a 0 means you were not tired at all and a 6 means you were very tired, how tired did you feel during this time?
3. Stressed  From 0 – 6, where a 0 means you were not stressed at all and a 6 means you were very stressed, how stressed did you feel during this time?
4. Sad  From 0 – 6, where a 0 means you were not sad at all and a 6 means you were very sad, how sad did you feel during this time?
5. Pain  From 0 – 6, where a 0 means you did not feel any pain at all and a 6 means you were in severe pain, how much pain did you feel during this time if any?
6. Meaningful  From 0 to 6, how meaningful did you consider what you were doing? 0 means it was not meaningful at all to you and a 6 means it was very meaningful to you.

[THE ORDER OF THE AFFECTIVE DIMENSIONS (ITEMS 1-5) WAS RANDOMIZED BY RESPONDENT].

7. Were you interacting with anyone during this time, including over the phone? (Yes/No)

The next set of questions asks about the respondent’s health in general.

PAIN ITEM # 1
10. Did you take any pain medication yesterday, such as Aspirin, Ibuprofen or prescription pain medication?
[INTERVIEWER NOTE: IF MENTIONS A DRUG, CODE AS A YES. FOR EXAMPLE, TYLENOL AND ALEVE ARE BOTH PAIN MEDS.]
☐ ☐ Yes
☐ ☐ No
HEALTH STATUS # 1
11. Finally, I have a couple of questions about your health. Would you say your health in general is excellent, very good, good, fair, or poor?
1. EXCELLENT
2. VERY GOOD
3. GOOD
4. FAIR
5. POOR
9. DON'T KNOW/REFUSED

HEALTH STATUS # 2
12. In the last five years, were you ever told by a doctor or other health professional that you have hypertension, also called high blood pressure, or borderline hypertension?
☐ ☐ Yes
☐ ☐ No

HEALTH STATUS # 3
13. When you woke up yesterday, how well-rested did you feel? Did you feel very rested, somewhat rested, a little rested, or not at all rested?
☐ ☐ Very
☐ ☐ Somewhat
☐ ☐ A little
☐ ☐ Not at all

In 2012 and 2013 two questions about general wellbeing were asked that unfortunately did not appear in the 2010 Wellbeing Module:

8. [CANTRIL] Please imagine a ladder with steps numbered from zero at the bottom to ten at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you.

If the top step is 10 and the bottom step is 0, on which step of the ladder do you feel you personally stand at the present time?

9. [TYPICAL] Thinking about yesterday as a whole, how would you say that your feelings, both good and bad, compared to a typical [fill day of the week]? Were they better than a typical [fill day of the week], the same as a typical [fill day of the week], or worse than a typical [fill day of the week]?
Data Quality in the 2010 ATUS

Even though time-use diaries have been established as the most advantageous method to measure time use (Juster, 1985a), especially because forgetting is minimized due to the shortness of the reference period and because respondents are only required to report activities from the previous day (Al Baghal, Belli, Phillips, & Ruther, 2014), there is still room for recall error. For instance, Al Baghal et al., 2014 point out that in the 2010 ATUS at least 4.64% of respondents have an activity recall failure (memory gap). Besides the ATUS specified errors, there are other indicators of data quality. For instance, there can be item non-response related to the place where the activity occurred and/or the person with whom the respondent was at that moment. Additionally, satisficing can be indicated by instances of rounding, in which the respondent does not make an effort to accurately provide an answer of how much time was spent on a particular activity and gives a rough rounded estimate (Fricker, 2007).

There is a score of papers and articles using the ATUS. However, most of them focus on how Americans use their time, and do not deal with issues of data quality. In fact, to the best of my knowledge, there are only three published articles specifically dealing with data quality in the ATUS and a few unpublished manuscripts and presentations on this topic. The published pieces include (1) Fricker and Tourangeau (2010), which is based on his dissertation (Fricker, 2007), (2) the book chapter by Phipps and Vernon (2009)\(^\text{10}\), and (3) Al Baghal, et al. (2014).

\(^{10}\) An important unpublished piece is that of Bose and Sharp (2005), which compares travel estimates between the 2001 National Household Travel Survey and the 2003 American Time Use Survey, and were error stemming from rounding was highlighted.
**Study examining the relationship between response propensity and data quality in the ATUS.** Fricker and Tourangeau (2010) examine the relationship between response propensity and data quality in the ATUS and the CPS. They were able to show that data quality decreased as the probability of nonresponse increased. The strength of such relationship, however, varied by the specific data quality indicator and survey (ATUS or CPS). According to these authors, one mechanism that will produce high quality responses is that of social capital. Broadly construed, social capital refers those “networks together with shared norms, values and understandings that facilitate cooperation within or among groups” (Keeley, 2007, p. 103). Thus, social capital can be understood as links between groups or individuals, including networks of friends, family members, work colleagues, etc. As Keely (2007) argues, “these networks and understandings engender trust and so enable people to work together” (p.103). In line with this definition, for Fricker and Tourangeau (2010): “higher levels of social capital could activate stronger norms of cooperation (producing higher response propensities) and those same norms also could influence respondents’ willingness to engage in more careful processing of the survey questions” (p. 936). Their indicators of social capital in the ATUS include: number of non-family/relatives present, employment status, marital status, presence of young children, median family income, and racial diversity.

A second mechanism influencing the quality of the responses is what Fricker and Tourangeau call “busyness or time-stress”. According to them, this mechanism could “produce a general disinclination both to participate in surveys and to respond accurately if interviewed” (p. 936). They use two variables as indicators of this construct in the
ATUS: percent of household adults who work, and occupation type (executive/professional, service, support/production, not in labor force). Finally, a third mechanism impacting data quality is that of “survey burden”. This mechanism is believed to have an effect because the ATUS respondents come from the CPS pool of respondents who have already been answering questions for eight waves of the CPS survey. It is argued that after responding to many requests, the quality of the data provided might be lower. In effect, according to Fricker and Tourangeau (2010), to the extent that individuals’ response propensities are positively correlated with the level of effort that is engaged during the response process, converting reluctant respondents might increase measurement error and reduce the quality of estimates (p. 935). Specifically, the more reluctant the respondent was to participate in the CPS and in the ATUS (measured through the paradata obtained from the Contact History Data), the lower the quality of the estimates.

Fricker and Tourangeau examine data quality not from a total survey error integrated perspective, as each indicator of quality (total number of activities, round durations, missing diary reports and labor force item nonresponse) was evaluated separately. The association of nonresponse propensity and each of the separate indicators of data quality was then examined. My dissertation, however, aims at evaluating data quality from an integrated vantage point, as data quality is not a series of independent and unrelated indicators, but is a latent construct that can be observed through several indicators. In that sense, an index of data quality will be constructed using several observed variables, including the four data-quality indicators that were examined in
Fricker and Tourangeau (2010), and taking account the theoretical mechanisms by which high-quality responses are said to be obtained.

**Study examining recall failures in the ATUS.** Al Baghal et al. (2014) draw heavily on the literature of the structure of autobiographical memory and how it is reported, as well as the strategies to enhance memory retrieval (Belli, 1998; Belli et al., 2001). Their main assumptions of autobiographical memory include that “temporal linkages among adjacent events, along with thematic relationships, are known to structure autobiographical memory” (p. 521), and that there exist at least three types of cuing mechanisms that enhance memory recall: top-down, sequential and parallel. Further, they argue that these are some of the features that may explain why data from calendar interviewing may be of better quality than that of standardized interviews.

Al Baghal et al. examine the role of distinctiveness on data quality for which they draw on the notion that events are better remembered when they are distinctive from other events, a feature that allows one to locate certain episodes in a more reliable way (Burton & Blair, 1991; Menon, 1993). However, in their research, they do not focus on the impact of distinctiveness in reporting an activity, but on the impact of distinctiveness in reporting the next activity. They hypothesize that the amount of distinctive information of what is remembered from each previous activity will determine how effective is that cue to remember the immediately following adjacent activity.

In this dissertation I will not try to replicate Al Baghal et al. (2014), but I will borrow some of their notion in the construction of my indexes. First, the variable of recall error will be part of my index of data quality. Also, I will also a somewhat different
interviewer characteristic, namely the interviewer’s cooperation rate in the 2010 ATUS. Finally, I will use another variable used in their study, namely the timing of the event with respect to the time when the interview took place. They argue that the timing of the event can affect recall, in that periods at the beginning of the day are more distant in memory than those at the end of the day; importantly, since some respondents are surveyed at different times during the next day, it is hypothesized that those surveyed later in the day will have a longer recall period than those surveyed earlier in the day (p. 526). In my dissertation, this variable will be included in the interview rapport construct.

**Interviewer-Respondent Interaction and Rapport in the 2010 ATUS**

After presenting some theoretical perspectives on how to understand and evaluate the construct of interview rapport, I now turn to some empirical pieces also shed light on this construct. Firstly, I will refer to a piece by Freedman et al. (2013), and secondly I will again examine at Fricker & Tourangeau (2010).

Freedman et al. (2013) maintain that they are the first researchers to look at interviewer and respondent interactions to assess data quality in a time diary study. However, I argue that what they actually did was to assess the respondent-interviewer interaction in a time diary, and not data quality. In effect, through the use of behavior codes of the utterances of respondents and interviewers of the 2009 Disability and Use of Time Survey (DUST), they tried to systematically describe interviewer-respondent interactions and find their relationship with a measure of “perceived time diary quality” (p. 72). Some of their findings are relevant to the topic discussed in this section. For instance, in their behavior-code evaluation of the different entries, they found that most
time diary questions are answerable by respondents, that is, the interview topics do not seem to be involved and for the most part are “codeable”, an important aspect in the context of time diaries. Likewise, they also found that only about 15% of respondent utterances signaled potential issues with comprehension of the question, again pointing at the interaction between interviewer and respondent, not the resulting data. Another important confirmation already hypothesized by Belli and Callegaro (2009) is that time diary questions “elicit conversation, even when questions are largely scripted, the purpose of which appears to be to promote the flow of the interview” (Freedman et al., 2013, p. 72). The important implication of this finding is that longer than average response times per interview did not necessarily indicate respondent difficulty with diary questions, but the fact that a conversation took place, signaling perhaps, a higher level of rapport between the respondent and the interviewer. And finally, an additional relevant aspect of this work by Freedman et al. is that they were able to use paradata (in their case, behavior coding) to evaluate the flow of the conversation in a time diary interview, or, in terms of this dissertation, to evaluate the level of rapport that emerged during the course of the interview.

Paradata in the 2010 ATUS

The ATUS instrument is designed through the BLAISE software that automatically collects paradata or process data. The ATUS paradata used in this dissertation was derived from the original Blaise audit trails, and contains data that describes the interaction between interviewers and the CATI instrument while entering responses offered by the ATUS respondents. These audit trails consist of text data that
captures every key stroke produced throughout the interview. There are two types of keystrokes in the audit trail files, (1) Action keystrokes that represent an interaction between the interviewer and the ATUS instrument with either a particular field for an activity (e.g., the Activity Type field, the Who field, the Where field etc.), and (2) Prompts that represents an interaction with an error “prompt” dialog box. An error prompt is a message that appears in the screen for the interviewer to either take some action or ask for clarification in connection with the respondent’s answers. For instance, prompts exist that require the interviewer to confirm whether the report that the respondent did not eat during the entire day is accurate (“eat check” prompt), or that an activity or a detail is missing in the diary (“missing activity”, “missing who”, “missing where” prompts).

For interactions with fields, the beginning of an Action is generally recorded with an “Enter Field” line and ends with a “Leave Field” line in the audit trail. For interactions with prompts, an Action comes from any corresponding “Dialog” and “Action” lines in the audit trail. This structure permitted the extraction of paradata variables and the generation of a rectangular paradata dataset. (Eck, 2016). Interactions with activity fields can be done with two purposes, (a) inserting or editing information within the field, or (b) navigating through the instrument. For any given activity, the majority of actions are simply navigations that occur very quickly (e.g., a click of the mouse); in some occasions, this simple navigations will result in duplicate actions recorded in the audit trails, but that need to be discarded as they don’t provide any
substantial information (e.g., pressing the left arrow, and immediately pressing the right arrow on the keyboard) (Eck, 2016).
CHAPTER 5
METHODS

Participants

The data consist of survey responses from the 2010 American Time Use Survey (ATUS), in which 13260 respondents participated and were asked to report each activity in which they were involved the previous day, how long the activity lasted, who they were with and where it took place. The public release file that contains these data was obtained from the BLS website (http://www.bls.gov/tus/data.htm). Two additional datasets are used in this dissertation that were merged to the referred public release file. The first additional dataset consists of the paradata obtained from the original Blaise audit trails, which had been sanitized by the Census Bureau to remove any personal identifying information. Given this is a computer-assisted telephone interview where every interviewer keystroke is captured automatically, this dataset contains information about the interaction by interviewers with the CATI instrument while entering responses provided by respondents. Figure 5.1 shows a screenshot of the CATI instrument with which the ATUS data was collected. The third dataset is also a public release data set which includes the data from the 2010 ATUS wellbeing module, in which respondents were asked about how they felt during selected activities, as well as general health information.
The public release file based on the 2010 ATUS respondents contains 13260 observations. The dataset with the audit trails that was obtained from the US Census Bureau did not have information for 62 respondents, while 436 ATUS respondents did not respond to the wellbeing module. Thus, after deleting those cases with incomplete information, 12762 respondents were left in the final dataset.

**Analytical Strategy**

**Objectives 1 and 2:** Using structural equations with latent variables to study total survey error. To answer the research questions in Objectives 1 and 2, I used a Structural Equation Model (SEM) to explore the relationship between the level of rapport
during the respondent-interviewer interaction and the resulting data quality (understood as the degree of completeness with which the survey was completed), while controlling for other important predictor variables. Thus, through the use of SEM, the purpose is to identify whether these two latent factors (i.e., rapport and data quality) account for the variation and covariation among certain variables found in the 2010 ATUS datasets, and how are they related to one another (Brown, 2006). Importantly, the decision to use SEMs is based on their ability to handle latent variables, measurement error, and multiple indicators, together with the possibility they offer to conduct test of model to assess the correspondence between models and data, as well as to compare different models (Bollen, Tueller, & Oberski, 2013). Thus, based on theory and previous research, which argues that those communicative strategies that are developed to establish “rapport” (i.e., the attempt to create a positive and friendly conversational interaction), may lead to the production of survey results of better quality (Olson & Bilgen, 2011), I estimated a structural equation model in which interview rapport predicts data quality, while controlling for other variables considered to affect both constructs.

The first step to explore the relationship between the constructs of rapport and data quality was to estimate a measurement model for each of the latent constructs of interest. To do so, I estimated a confirmatory factor analysis for each construct in Mplus v. 7.11. All models were identified by setting the latent factor mean to 0 and the latent factor variance to 1, such that all item intercepts, item factor loadings, and item residual variances were estimated. Given that the data are observational, each item has different scales. Accordingly, items were rescaled and in some cases reversed prior to analysis
when needed, so that the higher values then would indicate greater levels of interview rapport and data quality for their respective items. The distributional properties of each of the variables measuring each of the constructs were evaluated to insure the correct estimators were used. Table 5.1 shows the descriptive information about the indicator variables used.

Table 5.1

Descriptive Statistics of Measures

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of activities</td>
<td>12762</td>
<td>19.49</td>
<td>8.10</td>
<td>5</td>
<td>82</td>
</tr>
<tr>
<td>At least one basic missing activity</td>
<td>12762</td>
<td>0.24</td>
<td>0.43</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Rounding rate per interview</td>
<td>12762</td>
<td>0.76</td>
<td>0.13</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>At least one <em>who</em> missing in the diary</td>
<td>12762</td>
<td>0.36</td>
<td>0.48</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>At least one <em>where</em> missing in the diary</td>
<td>12762</td>
<td>0.50</td>
<td>0.50</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Duration of the interview in minutes</td>
<td>12762</td>
<td>10.70</td>
<td>6.16</td>
<td>1</td>
<td>72</td>
</tr>
<tr>
<td>Number of reported secondary activities</td>
<td>12762</td>
<td>0.84</td>
<td>1.58</td>
<td>0</td>
<td>42</td>
</tr>
<tr>
<td>At least one coded ATUS error in the diary</td>
<td>12762</td>
<td>0.19</td>
<td>0.39</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Reliability rate from Wellbeing Module</td>
<td>12762</td>
<td>0.77</td>
<td>0.28</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Total number of entries per diary</td>
<td>12762</td>
<td>138.14</td>
<td>67.38</td>
<td>20</td>
<td>789</td>
</tr>
<tr>
<td>At least one insert activity in the diary</td>
<td>12762</td>
<td>0.21</td>
<td>0.41</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Total number of <em>activity</em> changes</td>
<td>12762</td>
<td>28.23</td>
<td>33.97</td>
<td>0</td>
<td>884</td>
</tr>
<tr>
<td>Total number of <em>where</em> changes</td>
<td>12762</td>
<td>17.79</td>
<td>34.96</td>
<td>0</td>
<td>590</td>
</tr>
<tr>
<td>Total number of verbatim reports</td>
<td>12762</td>
<td>14.12</td>
<td>11.02</td>
<td>0</td>
<td>162</td>
</tr>
</tbody>
</table>
After estimating the measurement model for each of the constructs, following the recommended two-step approach for the estimation of SEM (Anderson & Gerbing, 1988; Kline, 2011), the first step before estimating the full structural model to predict data quality from interview rapport, was to re-specify the structural regression model (whose conceptual drawing is presented in Figure 5.2) as a confirmatory factor analysis measurement model. The CFA model was then analyzed in order to determine whether it fits the data (Kline, 2011). Once an acceptable CFA model was obtained, I proceeded to estimate the structural regression model.

Finally, it needs to be mentioned that given the intraclass correlation (ICC) coefficients for the observed variables used in the analyses, it was deemed necessary to account for the nesting of respondents within 69 interviewers. Table 5.2 shows that the ICCs for the observed variables used in the analyses range from 0.01 to 0.39. All of them are significant, and although some of them may appear to be small, if one takes into account that the average number of interviews per interviewer in the 2010 ATUS is about 192 (with one interviewer conducting up to 741 interviews), the impact of interviewer variance is not negligible. For instance, for a variable with an ICC of 0.01, the design effect can be as high as 2.70 for an interviewer with an average number of interviews.
Table 5.2

*Intraclass correlation coefficients for observed variables*

<table>
<thead>
<tr>
<th>Variable</th>
<th>ICC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of activities</td>
<td>0.04*</td>
</tr>
<tr>
<td>Interview duration</td>
<td>0.28*</td>
</tr>
<tr>
<td>Missing basic activity</td>
<td>0.03*</td>
</tr>
<tr>
<td>Missing <em>Who</em></td>
<td>0.12*</td>
</tr>
<tr>
<td>Missing <em>Where</em></td>
<td>0.39*</td>
</tr>
<tr>
<td>ATUS error</td>
<td>0.18*</td>
</tr>
<tr>
<td>Reliability from Wellbeing Module</td>
<td>0.03*</td>
</tr>
<tr>
<td>Rounding rate</td>
<td>0.01*</td>
</tr>
<tr>
<td>Number of verbatim reports</td>
<td>0.12*</td>
</tr>
<tr>
<td>Number of secondary activities reported</td>
<td>0.09*</td>
</tr>
<tr>
<td>Total number of entries in the diary</td>
<td>0.14*</td>
</tr>
<tr>
<td>Number of <em>where</em> changes</td>
<td>0.22*</td>
</tr>
<tr>
<td>Number of activity report changes</td>
<td>0.14*</td>
</tr>
<tr>
<td>Overall health reported</td>
<td>0.01*</td>
</tr>
</tbody>
</table>

* *p < .05

**Data Quality Measurement Model.** Following the application of social exchange theory in surveys (Dillman, Smyth, & Christian, 2014), that posits that under certain circumstances respondents will be more willing to cooperate with a survey task (e.g., when they are given a material incentive, or legitimately believe their opinion will benefit themselves and society as a whole), and the theory of errors in surveys, I consider two sources of error may threaten the degree of completeness when responding to the ATUS: measurement and nonresponse error. Through the estimation of a measurement model for data quality, understood as the degree of completeness of the reports about the respondent’s activities during the diary day, I explored the joint contribution of each of these error sources. The chosen observed indicator variables of the completeness of the
reports are theoretically linked to the ability and motivation to respond correctly. For instance, two of the previously most studied indicators of data quality in time diaries, namely the total number of reported activities and the non-failure to report a basic activity such as eating or sleeping, is not necessarily an indicator of the respondent’s day per se, but it can be assumed to be a reflection of the respondent’s willingness to respond to the diary in a thorough manner, together with the interviewer’s ability to efficiently probe for the required information. In that sense, one can contend that this measurement model is actually a reflective model (i.e., one in which the chosen indicators can be taken to be the effect of the factor), and not a formative model (i.e., one in which the indicators cause the factor). In effect, I argue that it is the latent factor of the propensity to provide as thorough and complete reports as possible what produces the resulting number of reported activities or the complete reports in terms of the additional details of the activity (was someone else present, and where the activity took place). For theoretical purposes, I have grouped the indicators of data quality into indicators of measurement and non-response error, but in the model they were tested simultaneously.

*Indicators of measurement error.*

1) Total number of diary activities reported: based on results from Fricker and Tourangeau (2010)\(^{11}\); it is hypothesized that a higher number of reported activities reflect a higher motivation to respond, and thus better data quality.

\(^{11}\) Fewer activities were reported by respondents with high nonresponse propensities and for those without children; the number of activity reports also was negatively correlated with hours worked and positively correlated with educational attainment (Fricker and Tourangeau, 2010, p. 949)
The average number of reported total activities was 19.5. This measure comes from the ATUS public release dataset.

2) Missing diary reports of basic daily activities: Following Fricker and Tourangeau (2010), I acknowledged that most people should sleep, eat, and engage in personal-care activities (e.g., grooming, dressing, etc.) on a daily basis. If the 24-hour diary does not contain one or more of these basic activities, it suggests that respondents either intentionally omitted some behaviors or failed to report their activities accurately (p. 942). Because more missing diary activities suggest lower data quality, this item is reversed so that higher values for this item indicate greater levels of data quality. Given the small proportion of failing to report to sleep (0.13%) and eating (4.07%), these two types of missing activities were added together to failing to report grooming (24.40%). Thus, this variable indicates whether any basic activity was missing. This measure comes from the ATUS public release dataset.

3) Round values for activity durations: Following Fricker (2007), I consider that a potential source of reporting error is rounded activity durations. Rounding is thought to reflect data quality issues in that “round value reports can occur because respondents misinterpret the question’s intent (i.e., the level of precision required), lack precise knowledge of the characteristic of interest, or are not motivated to provide a fully accurate answer. In providing an imprecise yet ‘plausible’ answer, respondents have a systematic tendency to report prototypical values (often multiples of 5 or 10)” (Fricker, 2007, p. 93).
To create this variable, any activity that was reported as lasting either 5, 10, 15, 20, 30, or 45 minutes, or 1, 1.5, 2, 2.5, 3, 3.5, 4, 5, 6, or 7 hours was marked as being a rounded report. A higher number of round values are an indication of lower data quality, so this item was reversed. A rounding rate was created per interview, where the number of total rounded reports were divided by number of total activities. The average rounding rate is 0.77. This measure was created from the ATUS public release dataset.

4) Reliability Scale: Taking advantage that in the Wellbeing module (where three activities are randomly selected from the pool of all previously reported activities), the respondent is asked a second time whether somebody else was present during the selected activity, a reliability scale was created by comparing the answers to whether somebody else was present from the actual diary where the activity was reported and the wellbeing module. The scale was constructed by summing the answers that were consistent on the two sections of the interview. Since the question of whether somebody else was present is asked twice in the same survey, this measure provides an indication of test-retest reliability. This measure comes from the ATUS public release dataset and the ATUS wellbeing module dataset.

5) Interview duration: Indicates the total time taken to complete an interview in minutes. Following Freedman (2007), I assumed that the longer the time taken to complete the interview indicates a higher motivation to provide thorough and complete answers to the diary. One may assume that more time spent per
interview means that the interviewer-respondent interaction was more fruitful and that respondents provided a more complete report of the details. This indicator might also mean there were cognitive issues on the part of the respondent or that there were problems in the interaction between the interviewer and the instrument, and that was the reason why the interview took a longer time. But even in that case, one may assume that the there was a higher motivation to provide more accurate responses as respondents were not speeding or satisficing. This measure comes from the audit trail dataset.

*Indicators of item nonresponse error.*

6) Diary ATUS errors: The ATUS has 5 types of “unable to code” errors. These errors mean that the interviewer was not able to categorize the reported activity within any of the existing pre-codes of the instrument to capture the respondent’s answer. These errors have a very low prevalence (1.44% from all activities reported by all respondents) and they include:

a. Insufficient Detail Error: The ATUS requires interviewers to probe for additional detail for certain activities so that they can be correctly coded in the post-collection phase (e.g., interviewer should ask whether a reported reading activity was done for work, class, volunteering, personal interest or another purpose) (0.80%)

---

12 According to the ATUS documentation, Missing Travel and Insufficient detail errors “directly affect ATUS data quality because the activity information will be missing. Census staff track the number of activities assigned these data codes on a quarterly basis. Interviewers with unacceptable rates of interviewer error codes receive additional guidance and training. Those with high error rates are taken off of ATUS interviewing until they pass a re-qualification test” (U.S. Bureau of Labor Statistics, 2014b).
b. Missing Travel Error: The ATUS error by which a failure to provide information about commuting. This indicates that interviewer failed to probe enough or that the respondent refused to provide information about the additional detail for commuting related activities (0.19%)

c. Simultaneous Activities Error: The main and the secondary activity were recorded incorrectly (0.14%)

d. Respondent refused to provide information (0.02%)

e. Diary memory gaps: number of instances in which the respondent could not report an activity carried out during the previous day such that there is a memory gap in his/her time diary. This is what Al Baghal et al. (2014) call “Recall failure” (0.26%)

f. Uncodeable Activity Error: The activity was unable to be coded at the first tier (0.03%).

Because of the low prevalence of each of these errors, all of the errors were added up per respondent. The final error variable used in the analyses indicates whether there was at least one of these errors in the person’s diary. There were 18.86% of respondents who at least have one such error in their diary. Because having an error means lower data quality, this variable was reversed for the analysis. This measure comes from the public release dataset.

7) Instances in which the respondent failed to report the place of the occurrence of the activity. Using the audit trails it was found that 49.84% of respondents
had at least one “missing where” code while responding to their diaries. Because failing to provide additional details such as “where” may mean a lack of the precise knowledge of the characteristic of interest, or a lack of motivation to provide a fully accurate answer, it is considered a sign of lower data quality, so this variable was reversed for the analysis. This is a ‘process quality’ variable, as it not necessarily reflects the end result (a final missing where code). This variable was created from the ATUS paradata dataset.

8) Instances in which the respondent failed to report whether somebody else was present during the occurrence of the activity. Using the audit trails it was found that 36.11% of respondents had at least one “missing who” error in their diaries. Because failing to provide additional details such as “who” means lower data quality, this variable was reversed for the analysis. This is also a ‘process quality’ variable, as it not necessarily reflects the end result (a final missing who code).

Table 5.2 shows that the item-total correlations are moderate to low, and in some cases negative. It seems that the items most strongly related to the latent construct of data quality are total number of activities reported and duration of the interview. The Alpha coefficient is also moderate (.33). Table 5.3 contains the Cronbach’s Alpha results and the correlations between the variables used to create the data quality index.
Table 5.3

*Cronbach Coefficient Alpha for Data Quality Indicators*

<table>
<thead>
<tr>
<th>Deleted Variable</th>
<th>Raw Variables</th>
<th></th>
<th>Standardized Variables</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correlation</td>
<td>Alpha</td>
<td>Correlation</td>
<td>Alpha</td>
</tr>
<tr>
<td></td>
<td>with Total</td>
<td></td>
<td>with Total</td>
<td></td>
</tr>
<tr>
<td>Total number of activities reported in the diary</td>
<td>0.698</td>
<td>0.072</td>
<td>0.373</td>
<td>0.166</td>
</tr>
<tr>
<td>At least one basic missing activity in the diary</td>
<td>-.283</td>
<td>0.502</td>
<td>-.190</td>
<td>0.463</td>
</tr>
<tr>
<td>Rounding rate per interview</td>
<td>-.007</td>
<td>0.490</td>
<td>-.007</td>
<td>0.377</td>
</tr>
<tr>
<td>At least one “who” missing in the diary</td>
<td>0.266</td>
<td>0.477</td>
<td>0.235</td>
<td>0.249</td>
</tr>
<tr>
<td>At least one “where” missing in the diary</td>
<td>0.241</td>
<td>0.478</td>
<td>0.198</td>
<td>0.270</td>
</tr>
<tr>
<td>Duration of the interview in minutes</td>
<td>0.706</td>
<td>0.051</td>
<td>0.409</td>
<td>0.144</td>
</tr>
<tr>
<td>At least one coded error in the diary</td>
<td>0.223</td>
<td>0.481</td>
<td>0.173</td>
<td>0.283</td>
</tr>
<tr>
<td>Sum of 'reliable' items from Wellbeing Module</td>
<td>0.002</td>
<td>0.491</td>
<td>-.024</td>
<td>0.386</td>
</tr>
</tbody>
</table>

Note. N=12762
Table 5.4

*Pearson Correlation Coefficients for Data Quality Indicators*

<table>
<thead>
<tr>
<th></th>
<th>Total number of activities</th>
<th>Basic activity missing</th>
<th>Rounding rate per interview</th>
<th>At least one “who” missing</th>
<th>At least one “where” missing</th>
<th>Duration of the interview</th>
<th>At least one coded error</th>
<th>Sum of reliable items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of activities</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic activity missing</td>
<td>-0.30*</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rounding rate per interview</td>
<td>0.01</td>
<td>0.04*</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At least one “who” missing</td>
<td>0.24*</td>
<td>-0.07*</td>
<td>-0.02</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At least one “where” missing</td>
<td>0.21*</td>
<td>-0.06*</td>
<td>-0.04*</td>
<td>0.24*</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration of the interview</td>
<td>0.70*</td>
<td>-0.21*</td>
<td>-0.03</td>
<td>0.24*</td>
<td>0.23*</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>At least one coded error</td>
<td>0.17*</td>
<td>-0.03*</td>
<td>0.01</td>
<td>0.08*</td>
<td>0.05*</td>
<td>0.25*</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Sum of reliable items</td>
<td>0.03*</td>
<td>-0.02</td>
<td>0.01</td>
<td>-0.01</td>
<td>-0.04*</td>
<td>-0.03*</td>
<td>-0.01</td>
<td>1.00</td>
</tr>
</tbody>
</table>

*Note. N = 12762
* p < .05.
Index of Interview Rapport. Taking into account the concept of interview rapport, understood as the strategy whereby an interviewer creates a friendly environment with the aim of enabling a more productive and cooperative survey interaction, I will estimate a measurement model for an interview rapport factor, which I believe is the underlying cause for the covariation between the observed variables of changes in previous reports, the number of entries per activity, the number of secondary activities reported, and the number of verbatim reports. I believe that rapport is what causes the covariation between those variables because these indicators seem to signal that a more flexible and productive interaction took place during the interview; in effect, they seem to speak to the fact that the relationship that was established during the interview was friendly and harmonious enough to allow for changes in the already reported activities or the extra effort of reporting secondary activities, which are not really required in the ATUS. A description of the rapport indicators follows:

1) Total number of action entries per respondent: This is a paradata variable that indicates interactions between the interviewer and the ATUS instrument with regards to entering an action such as an activity, a “who”, or a “where” code. In principle, there should be only one entry per each of those actions; however, if there are more entries than actions, this can be taken as a signal of changes going on because of the interaction that is taking place. In this dissertation, I assume that there are more entries because, thanks to the level of rapport or reciprocity between the respondent and the interviewer, there is room for changes and “repairs” in the conversation. A higher number of
entries are taken to reflect a higher level of interaction and rapport. This variable comes from the audit trails dataset.

2) Total number of times in which the respondent reported a secondary activity: Secondary activities are those activities that are done simultaneously with the main activity reported. Given that the ATUS diary interview is mainly concerned with the primary activity and that interviewers are not even asked to probe for secondary activities, the reporting of a secondary activity can be taken as an indicator of a higher rapport during the interaction that allows for a richer interaction to take place.

3) Number of ‘activity’ edits: This paradata variable records changes in an activity that had already been recorded. This variable is taken to reflect a richer interaction and higher flexibility during the interview, as changes in a previously reported activity indicates the question-answer process had to be ‘repaired’ in some manner, and the ‘friendliness’ between interviewer and the respondent allowed for that. A higher number of activity edits is taken to reflect more interview rapport. This variable comes from the audit trails dataset.

4) Number of ‘where’ edits: This paradata variable records changes in the place of occurrence of an activity that had already been recorded. This variable is taken to reflect a richer interaction in the interview as changes in a previously reported ‘where’ code indicates the question-answer process had to be ‘repaired’ in some manner, and the interaction was friendly enough to allow
for that change. A higher number of where changes is taken to reflect more interview rapport. This variable comes from the audit trails dataset.

5) Number of verbatim reports per interview: of instances in which the respondent’s answer was recorded verbatim, that is, a predefined code was not used to record the activity. This is taken to be an indicator of a richer interaction between the respondent and the interviewer, as it is taken to signal that rather than stopping the flow of the conversation, the interviewer preferred to write down the response (as quickly as possible) A higher number of verbatim reports is an indicator of higher interview rapport. This variable comes from the audit trails dataset.

Tables 5.4 and 5.5 contain the Cronbach’s Alpha results and the correlations between the variables used to create the interview rapport index. Table 5.5 shows that the item-total correlations were moderate to high. The total number of entries and number of activities reported verbatim are the items most strongly related to the latent construct of interview rapport. The Alpha coefficient for these items was high (.73).
### Table 5.5

**Cronbach Coefficient Alpha for Interview Rapport Indicators**

<table>
<thead>
<tr>
<th>Deleted Variable</th>
<th>Raw Variables</th>
<th>Standardized Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correlation with Total</td>
<td>Alpha</td>
</tr>
<tr>
<td>Number of entries</td>
<td>0.87</td>
<td>0.51</td>
</tr>
<tr>
<td>Number of secondary activities</td>
<td>0.37</td>
<td>0.76</td>
</tr>
<tr>
<td>Number of activity changes</td>
<td>0.56</td>
<td>0.70</td>
</tr>
<tr>
<td>Total of <em>where</em> changes</td>
<td>0.36</td>
<td>0.74</td>
</tr>
<tr>
<td>Number of verbatim reports</td>
<td>0.84</td>
<td>0.61</td>
</tr>
</tbody>
</table>

*Note. N = 12762*

### Table 5.6

**Pearson Correlation Coefficients for Interview Rapport Indicators**

<table>
<thead>
<tr>
<th></th>
<th>Number of entries</th>
<th>Number of secondary activities</th>
<th>Number of activity changes</th>
<th>Number of “where” changes</th>
<th>Number of verbatim reports</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of entries</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of secondary activities</td>
<td>0.31*</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of activity changes</td>
<td>0.53*</td>
<td>0.50*</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total of <em>where</em> changes</td>
<td>0.37*</td>
<td>0.06*</td>
<td>0.24*</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Number of verbatim reports</td>
<td>0.90*</td>
<td>0.35*</td>
<td>0.53*</td>
<td>0.35*</td>
<td>1.00</td>
</tr>
</tbody>
</table>

*Note. N = 12762*

*p < .05.*
**Control variables.** The control variables hypothesized to impact data quality and rapport include:

1) Respondent demographics: Age, gender, and education.

2) Timing of the interview: Given that the timing of event can affect recall, it is hypothesized that those surveyed later in the day will have a longer recall period than those surveyed earlier in the day.

3) Interviewer’s cooperation rate in the 2010 ATUS.

4) Busyness\(^{13}\): Following Fricker and Tourangeau (2010), busyness will be indicated by the percent of household adults who work.

5) Social capital variables as presented in Fricker and Tourangeau (2010) including the number of non-family/relatives present, employment status, marital status, presence of young children, median family income, and racial diversity.

6) Participation Reluctance: Reluctance will be measured by the total number of weeks that calls were attempted before the interview was completed.

---

\(^{13}\) Busyness is understood here as the quality of being busy.
Figure 5.2 Conceptual Drawing of SEM Model to be Estimated
**Objective 3: Wellbeing “fidelity” analysis.** In this analysis a multiple regression model that predicts overall health taken as a proxy of overall wellbeing was estimated. I closely follow Juster et al. (1985) (See Chapter 2 above) and Krueger et al. (2009) in the selection of predictor variables of wellbeing. These include: income, family composition, employment status, intangible capital stocks, and measures that evaluate the respondent’s use of time. The objective of this analysis was to disattenuate the relation between wellbeing and its predictors by controlling for the quality of the data and the level of rapport.

To control for data quality and interview rapport, I used the factor scores previously constructed from their confirmatory factor analysis measurement models. These factor scores can be understood as imputed values for latent variables, or what would have been observed if there were a way to measure data quality in a direct manner (Muthén & Muthén, 2010). In other words, they allow to determine a participant’s relative standing on the latent dimension (Brown, 2006). The factor scores for each latent construct were created using the SAVE = FScores option in Mplus v. 7.11. This command creates factor scores using a frequentist estimation, and these are available when observed dependent variables are continuous, censored, binary, ordinal, or count (Muthén & Muthén, 2010, p. 756). Finally, these factor scores were treated as survey weights in the referred regression model. I converted into positive numbers all the negative factor scores by adding a constant (whose value was higher to the most extreme negative value in each factor) to all observations. This multiple regression analysis was conducted in SAS 9.4 using PROC SURVEYREG, accounting for the nesting within
interviewers by using the CLUSTER option, and conducting separate analyses with the each of the different factor scores weights by using the WEIGHT option.

It was expected that the inclusion of the data quality and interview rapport weights would contribute to refine the interpretation of results. Specifically, the purpose of these analysis was to evaluate the “fidelity” of the measures of wellbeing vis-à-vis the quality of the time-use data and the rapport with which the survey was conducted. In this sense, ‘fidelity’ is understood as the capacity of a method to provide more accurate answers about a narrow question, often at the cost of “bandwidth” (Shadish, Cook, & Campbell, 2002, p. 507), which refers to the capacity of a method to answer a broad array of questions “even if the causal question is answered less well” (Shadish et al, 2002, p. 98). Thus, with this analysis, I tested whether it is possible to “fine-tune” the wellbeing estimates by assessing whether the quality and rapport weights modified in some way the relationship between wellbeing and its predictors.

**Variables used in the analysis.** Following Juster et al. (1985) and Krueger et al (2009) the following are the variables that predict wellbeing:

1) Average household income, as a measure of tangible resources

2) Living arrangements, as a measure of tangible assets. This variable indicates whether the respondent’s living quarters are owned or rented

3) Marital status, as a measure of the respondent’s organizational context

4) Employment status, as a measure of the respondent’s organizational context

5) Number of children, as a measure of the household production process
6) Activity-based wellbeing measures taken from the answers to the questions in the wellbeing module where respondents were asked how they felt the day before doing specific activities. These are measures that evaluate the respondent’s use of time and serve as subjective measures of satisfaction connected to the activities of individuals. For each of the three activities selected, respondents needed to answer whether they felt the activity was meaningful, or if they felt happy, sad, tired, stressed, or in pain. The mean for the affect responses for all activities was obtained and included in the analysis (i.e., the mean of ‘meaningful’, ‘happy’, ‘sad’, ‘tired’, ‘stressed’, ‘in pain’ for all three selected activities).
CHAPTER 6
RESULTS

Structural Model Predicting Data Quality from Interview Rapport

Measurement model for data quality. The reliability and dimensionality of eight items each purportedly tapping into the construct of data quality was assessed in a sample of 12,762 respondents to the 2010 ATUS with a confirmatory factor analysis using Mean and Variance adjusted Weighted Least Square (WLSMV) estimator and Theta scaling in Mplus v. 7.11. The WLSMV is used because of the presence of categorical and count variables, and because model results provide goodness-of-fit indices necessary to compare competing models.

All models were identified by setting the latent factor mean to 0 and the latent factor variance to 1, such that all thresholds and item factor loadings estimated. Additionally, the CLUSTER option was used under TYPE=COMPLEX to indicate that respondents are nested within 69 interviewers in this dataset. As previously indicated, because of non-negligible ICCs among the indicator variables (See Table 5.2), such nesting needs to be taken into consideration. Because of extreme observations, some variables were truncated at the 99 percentile. The truncated variables include total number of activities and interview duration in minutes. Figures 6.1 to 6.8 show the distribution of all the data quality indicators.
Figure 6.1. Distribution of total number of activities reported in the diary (Truncated at the 99 percentile).

Figure 6.2. Distribution of whether at least one basic activity was missing in the diary.
Figure 6.3. Distribution of whether at least one ‘who’ was missing in the diary.

Figure 6.4. Distribution of whether at least one ‘where’ was missing in the diary.
Figure 6.5. Distribution of interview duration in minutes (Truncated at the 99 percentile).

Figure 6.6. Distribution of whether there was at least one ATUS error in the diary.
Figure 6.7. Distribution of rounding rate.

Figure 6.8. Distribution of number of consistent (“reliable”) items from the comparison of response to diary and wellbeing module.
Given this is an observational study, each item had different scales and not all of the indicators had a normal distribution. Some items were reversed so that the higher values would then indicate greater levels of data quality for all items. Specifically, prior to analysis, the indicators of missing a basic activity report, rounding rate, missing who, missing where, and error in diary were reversed.

Model fit statistics reported in Table 6.1 include the obtained model $\chi^2$, its degrees of freedom, and its $p$-value (in which non-significance is desirable for good fit, although in this case significance will be expected due to the large sample size); CFI, or Comparative Fix Index (in which values higher than .95 are desirable for good fit), and the RMSEA, or Root Mean Square Error of Approximation, point estimate and 90% confidence interval (in which values lower than .06 are desirable for good fit) (Hu & Bentler, 1999). Nested model comparisons, shown in Table 6.2, were conducted the $\chi^2$ Test for Difference Testing implemented by Mplus using the DIFFTEST option. Following Cheung and Rensvold (2002), who argue that if the sample sizes are large, even a small difference may result in a significant value of $\Delta \chi^2$, the cut-off level of significance for these comparisons is $p<.01$. The specific models examined are described in detail below.
### Table 6.1
Assessment of Model Fit using WLSMV for Data Quality Measurement Model

<table>
<thead>
<tr>
<th>Model</th>
<th># Items</th>
<th>$\chi^2$</th>
<th>$\chi^2$ DF</th>
<th>p-value</th>
<th>CFI</th>
<th>RMSEA</th>
<th>RMSEA Lower CI</th>
<th>RMSEA Higher CI</th>
<th>RMSEA p-value</th>
<th>RMSEA A p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-Factor</td>
<td>8</td>
<td>216.213</td>
<td>20</td>
<td>&lt;.0001</td>
<td>.907</td>
<td>.028</td>
<td>.024</td>
<td>.031</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>One-Factor (Rounding rate and reliability removed)</td>
<td>6</td>
<td>199.511</td>
<td>9</td>
<td>&lt;.0001</td>
<td>.911</td>
<td>.041</td>
<td>.036</td>
<td>.046</td>
<td>.999</td>
<td></td>
</tr>
<tr>
<td>Two factors: negative items set to load on a second factor</td>
<td>6</td>
<td>142.462</td>
<td>8</td>
<td>&lt;.0001</td>
<td>.937</td>
<td>.036</td>
<td>.031</td>
<td>.042</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Two factors, where error with who error</td>
<td>6</td>
<td>99.867</td>
<td>7</td>
<td>&lt;.0001</td>
<td>.956</td>
<td>.032</td>
<td>.027</td>
<td>.038</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Two factors, missing activities with number of activities</td>
<td>6</td>
<td>25.383</td>
<td>6</td>
<td>&lt;.0001</td>
<td>.991</td>
<td>.016</td>
<td>.010</td>
<td>.023</td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* WLSMV=Weighted Least Squares; CFI=Comparative Fit Index; RMSEA=Root mean square error of approximation.

### Table 6.2
Nested Model Comparisons for Data Quality Measurement Model

<table>
<thead>
<tr>
<th>Model</th>
<th>DF Difference</th>
<th>$\chi^2$ Test Difference</th>
<th>$\chi^2$ Test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA QUALITY</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One-Factor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two factors: negative items set to load on a second factor</td>
<td>1</td>
<td>57.239</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Two factors, <em>where</em> error with <em>who</em> error</td>
<td>1</td>
<td>37.318</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Two factors, basic activities with number of activities</td>
<td>1</td>
<td>97.282</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

*Note.* DF=Degrees of Freedom
Table 6.3

*Weighted Least Squares Estimates of Factor Loadings for a Data Quality Measurement Model*

<table>
<thead>
<tr>
<th>Model Parameter</th>
<th>Unstandardized Estimate</th>
<th>SE</th>
<th>Standardized Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Motivation</strong> Factor loadings</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of activities reported</td>
<td>6.055</td>
<td>0.162</td>
<td>0.771</td>
</tr>
<tr>
<td>Not missing a basic activity</td>
<td>0.361</td>
<td>0.024</td>
<td>0.339</td>
</tr>
<tr>
<td>Duration of the interview in minutes</td>
<td>5.393</td>
<td>0.166</td>
<td>0.924</td>
</tr>
<tr>
<td><strong>Memory</strong> Factor loadings</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing where code</td>
<td>0.427</td>
<td>0.043</td>
<td>0.392</td>
</tr>
<tr>
<td>Missing who code</td>
<td>0.376</td>
<td>0.037</td>
<td>0.352</td>
</tr>
<tr>
<td>ATUS error code</td>
<td>0.355</td>
<td>0.038</td>
<td>0.334</td>
</tr>
<tr>
<td><strong>Motivation WITH Memory</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing <em>where</em> WITH Missing <em>who</em></td>
<td>-0.939</td>
<td>0.073</td>
<td>-0.939</td>
</tr>
<tr>
<td>Not missing basic activities WITH Number of activities</td>
<td>1.695</td>
<td>0.13</td>
<td>0.339</td>
</tr>
<tr>
<td><strong>Intercepts</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of activities</td>
<td>19.416</td>
<td>0.299</td>
<td>2.473</td>
</tr>
<tr>
<td>Duration of the interview in minutes</td>
<td>10.630</td>
<td>0.439</td>
<td>1.822</td>
</tr>
<tr>
<td><strong>Thresholds</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not missing a basic activity$1$</td>
<td>-0.737</td>
<td>0.036</td>
<td>-0.693</td>
</tr>
<tr>
<td>No missing who code$1$</td>
<td>-0.386</td>
<td>0.06</td>
<td>-0.356</td>
</tr>
<tr>
<td>No missing where code$1$</td>
<td>-0.004</td>
<td>0.058</td>
<td>-0.004</td>
</tr>
<tr>
<td>No ATUS error code$1$</td>
<td>-0.937</td>
<td>0.07</td>
<td>-0.883</td>
</tr>
<tr>
<td>Total number of activities reported</td>
<td>24.988</td>
<td>1.742</td>
<td>0.405</td>
</tr>
<tr>
<td>Duration of the interview in minutes</td>
<td>4.963</td>
<td>1.267</td>
<td>0.146</td>
</tr>
<tr>
<td><strong>R-Square</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed Variable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of activities reported</td>
<td>0.595</td>
<td>0.027</td>
<td></td>
</tr>
<tr>
<td>Not missing a basic activity</td>
<td>0.115</td>
<td>0.014</td>
<td></td>
</tr>
<tr>
<td>No missing who code</td>
<td>0.154</td>
<td>0.026</td>
<td></td>
</tr>
<tr>
<td>No missing where code</td>
<td>0.124</td>
<td>0.021</td>
<td></td>
</tr>
<tr>
<td>Duration of the interview in minutes</td>
<td>0.854</td>
<td>0.037</td>
<td></td>
</tr>
<tr>
<td>No ATUS error code</td>
<td>0.112</td>
<td>0.021</td>
<td></td>
</tr>
</tbody>
</table>
The model initially posited to account for the pattern of covariance across the nine indicators did not fit the observed data exactly, \( \chi^2 (20) = 216.213, p<.05 \). However, a significant test of exact fit was expected given the large sample size. Likewise, based on the guidelines proposed by Hu and Bentler (1999), the fit of the one-factor was not acceptable by the test of approximate fit, as CFI=907, although there is evidence that the model may fit *reasonably* well in the population according to the error of approximation index (Brown, 1996), as RMSEA =0.028, \( p=1 \). Since effect sizes (standardized loadings) were very low for the rounding and the reliability scale (-0.022 and 0.002, respectively), and neither of the unstandardized loadings of these two indicators were statistically significant, I decided to remove these two indicators from the model. These results are not at all surprising given the original pattern where rounding rate and the reliability scale had very small item-total correlations. As it is shown in Table 6.1, when this was done, the fit of the model increased slightly, but was still not acceptable by the CFI criterion, CFI=.911. The model still have a reasonably good approximation, as RMSEA=.041, \( p=0.999 \); the test of exact fit still had to be rejected, \( \chi^2 (9) = 199.511, p<.05 \). Because this model is not nested in the previous one, the \( \chi^2 \) Test for Difference Testing could not been carried out.

Sources of local misfit were identified using the residuals for the estimated correlations. These showed several items with negative medium residual correlations. Additionally, I noted that three items that had been reversed prior to the analysis (missing *who* reports, missing *where* reports, and having an ATUS coded error in the diary) had negative loadings on the data quality factor. The need of two separate latent factors for
the negatively and positively loaded items was tested by specifying a two-factor model in which the negative items indicated data quality related to the non-response error indicators (not missing *who* reports, not missing *where* reports, and not having an ATUS error code), and the positive items seemed to indicate data quality related to the measurement error indicators (total number of reported activities, not missing a basic activity report, and the interview duration in minutes). The two factors were allowed to correlate. The two-factor model seemed to fit the data closely, as RMSEA=0.036, *p*=1, but still the test of approximate fit had to be rejected, as CFI=0.937. The model still did not have an exact fit to the observed data, as $\chi^2=142.462$, *p*<.05. In any case, the two-factor model fit significantly better than the one-factor model as indicated by the $\chi^2$ Test for Difference Testing that was significant ($\Delta\chi^2 (1) = 57.239$, *p*=.001). Therefore, the six indicators of data quality seem to measure two separate though negatively related constructs. Additionally, given that only three of the four indicators that were reversed are part of the second factor, and that conceptually they all belong to errors of non-response, it seems there is some evidence that the existence of this second factor is not due to an artifact of the difference between having positive and negative coded indicators. In fact, the significant negative correlation between the two factors (*r*= -0.655) seems to be an indication of two different processes governing each of the quality constructs: the missing *who*/*where* responses and ATUS error codes may be governed by memory failure, whereas the measurement quality construct may be governed by a process related to the respondent’s motivation to respond whereby satisficing is minimized.
Further examination of local fit via residual covariances and modification indices suggested that the ‘where’ errors and ‘who’ errors be correlated. Modification indices, available through the MODINDICES output option in Mplus corroborated this pattern. Given the structure of the diary interview, in which after reporting the activity one needs to report the place of occurrence and whether somebody else was present, it is theoretically justifiable to correlate the errors of these two items. When correlating these two variables, the model fit was found to be acceptable, as CFI=0.956 and RMSEA=0.032, p=1. Likewise, as indicated in table 6.2, the Test for Difference Testing was significant as well ($\Delta \chi^2 (1) = 37.318, p=.001$), so this model was retained. The test of exact fit was still significant due to the large sample size, $\chi^2=99.867, p<.05$. Finally, a large positive correlation between number of activities and the reversed variable for missing basic activities was identified (.390). Modification indices also suggested a relationship between these two variables. It is more likely that the greater number of reported activities, the lower the amount of missing basic activities, and therefore I decided to correlate the errors of these two variables. As Table 6.1 shows, although the $\chi^2$ remained significant, $\chi^2=25.383, p<.05$, the fit of the model further improved (CFI=0.991 and RMSEA=0.16, p=1). Further examination of local fit and modification indices did not yield additional interpretable relationships, so this two-factor model was retained.

Table 6.3 provides estimates and their standard errors for the item factor loadings and thresholds, from both the unstandardized and standardized solutions, as well as the R-squares for observed variables. All factor loadings and the factor covariance were statistically significant. As shown in Table 6.3, standardized loadings ranged from .334 to
.924 (with R² values for the amount of item variance accounted for by the factor ranging from .112 to .854), suggesting practical significant as well.

**Measurement model for interview rapport.** The reliability and dimensionality of six items each purportedly tapping into the construct of interview rapport was assessed in a sample of 12,762 respondents from the 2010 ATUS using a confirmatory factor analysis using robust maximum likelihood estimation (MLR) in Mplus v. 7.11. All models were identified by setting the latent factor mean to 0 and the latent factor variance to 1, such that all item intercepts, item factor loadings, and item residual variances were then estimated. Additionally, the CLUSTER option was used under TYPE=COMPLEX to indicate that respondents are nested within 69 interviewers in this dataset.

Additionally, as can be seen from figures 6.9 to 6.13, even though all indicator variables have a Poisson distribution, the only estimator that could be used to obtain fit indices with these distributions under the COMPLEX option was Robust Maximum Likelihood (MLR); WLSMV could not be used in this measurement model because of the lack of at least one categorical variables included in the model.

As was also needed with the data quality measurement model, some items were rescaled. Specifically due to very few extreme observations, prior to analysis, the variable for total number of entries and the number of ‘activity’ and ‘where’ changes were rescaled (grouped in groups of 10) and truncated at the 99 percentile (360, 150, 160, and 18 respectively); and the total number of secondary activities and verbatim reports were also truncated at the 99 percentile (42 and 55, respectively).
Figure 6.9. Distribution of total number of entries (Truncated at the 99 percentile).

Figure 6.10. Distribution of total number of secondary activities reported (Truncated at the 99.9 percentile).
Figure 6.11. Distribution of total number of changes in reported activities (Truncated at the 99 percentile).

Figure 6.12. Distribution of total number of ‘Where’ changes (Truncated at the 99 percentile).
Figure 6.13. Distribution of total number of verbatim reports (Truncated at the 99 percentile).

None of these variables needed to be reversed, as it is hypothesized that higher values indicate greater levels of interview rapport for all items. Model fit statistics reported in Table 6.4 include the obtained model $\chi^2$, its scaling factor (in which values different than 1.00 indicate deviations from normality), its degrees of freedom, and its $p$-value (in which non-significance is desirable for good fit), CFI, or Comparative Fix Index (in which values higher than .95 are desirable for good fit), and the RMSEA, or Root Mean Square Error of Approximation, point estimate and 90% confidence interval (in which values lower than .06 are desirable for good fit). Nested model comparisons, shown in Table 6.5, were conducted using the rescaled Log-likelihood difference test ($-2\Delta LL$) with degrees of freedom equal to the rescaled difference in the number of
parameters between models (i.e., a rescaled likelihood ratio test). Again, following Cheung and Rensvold (2002), and given the large number of observations, the cut-off level of significance for these comparisons is $p<.01$. The specific models examined are described in detail below.

The fit of the one-factor model initially posited to account for the pattern of covariance across the six indicators was not acceptable by any criterion (CFI=.905; RMSEA=0.063, $p=0.069$, and $\chi^2 (15)=312.412$), indicating that the model was not a close or approximate representation of the data. Further examination of local fit via normalized residuals indicated a large residual covariance between number of activity changes and number of secondary activities reported. The modification indices confirmed this pattern as well. Having previously examined the actual verbatim records, this relation is likely to have happened because the respondent decided to switch the originally reported main activity with the secondary activity, provoking a change in the activity field. As shown in Table 6.5, when this covariance was added, the rescaled difference was significant ($-2\Delta LL (1) = 105.892, p=.001$), so this model was retained because of its better fit. Indeed, when the additional covariance was added, there is evidence that the model adequately represents the data. According to the guidelines proposed by Hu and Bentler (1999), this model has a close approximate fit in the population, as CFI=0.981 and RMSEA=0.034, $p=1$. It is not possible to say this model fits the data exactly, as $\chi^2 (64.277), p<0.05$, but this was expected because of the large sample size. Further examination of local fit and modification indices did not yield additional interpretable
relationships, so this one-factor model with an estimated covariance between activity changes and number of secondary activities was retained.

The standardized loadings of this model range from .38 to .96 (with $R^2$ ranging from .14 to .92) suggesting moderate to high practical significance. Table 6.6 provides the estimates and their standard errors for the item factor loadings, intercepts, residual variances and R-Squares from both the unstandardized and standardized solutions. All factor loadings were statistically significant. Omega model-based reliability was calculated for this factor as described in Brown (2006) as the squared sum of the factor loadings divided by the squared sum of the factor loadings plus the sum of the error variances plus twice the sum of the error covariances. Omega was .91 suggesting high reliability for this five-item factor.
Table 6.4
Assessment of Model Fit using MLR for Interview Rapport Measurement Model

<table>
<thead>
<tr>
<th>Model</th>
<th># Items</th>
<th>$\chi^2$</th>
<th>Scaling Factor</th>
<th>$\chi^2$ DF</th>
<th>p-value</th>
<th>CFI</th>
<th>RMSEA</th>
<th>RMSEA Lower CI</th>
<th>RMSEA Higher CI</th>
<th>RMSE Ap-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-Factor</td>
<td>5</td>
<td>312.412</td>
<td>7.601</td>
<td>5</td>
<td>&lt;.0001</td>
<td>.905</td>
<td>.069</td>
<td>.063</td>
<td>.076</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>One-Factor (Activity changes with secondary activities)</td>
<td>5</td>
<td>64.277</td>
<td>4.591</td>
<td>4</td>
<td>&lt;.0001</td>
<td>.981</td>
<td>.034</td>
<td>.027</td>
<td>.042</td>
<td>1.000</td>
</tr>
</tbody>
</table>

*Note.* MLR=Robust Maximum Likelihood; CFI=Comparative Fit Index; RMSEA=Root mean square error of approximation.

Table 6.5
Nested Model Comparisons for Interview Rapport Measurement Model

<table>
<thead>
<tr>
<th>Model</th>
<th>H0 LL Scale Factor</th>
<th>Δ-2LL</th>
<th>Δ Scaling Correction</th>
<th>Scaled Δ-2LL</th>
<th>Δ DF</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-Factor</td>
<td>-164,824.604</td>
<td>35.226</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One-Factor (Activity changes with secondary activities)</td>
<td>-163,784.901</td>
<td>34.252</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test of Difference</td>
<td>-163,784.901</td>
<td>2,079.406</td>
<td>19.637</td>
<td>105.892</td>
<td>1</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

*Note.* DF=Degrees of Freedom; Δ-2LL=Minus 2 Log Likelihood Difference.
<table>
<thead>
<tr>
<th>Model Parameter</th>
<th>Unstandardized Estimate</th>
<th>SE</th>
<th>Standardized Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RAPPORT BY</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of Entries</td>
<td>6.108</td>
<td>0.205</td>
<td>0.954</td>
</tr>
<tr>
<td>Total number of Secondary Activities</td>
<td>0.533</td>
<td>0.054</td>
<td>0.342</td>
</tr>
<tr>
<td>Total number of Activity Changes</td>
<td>1.748</td>
<td>0.099</td>
<td>0.556</td>
</tr>
<tr>
<td>Total number of Where Changes</td>
<td>1.273</td>
<td>0.143</td>
<td>0.378</td>
</tr>
<tr>
<td>Total number of Verbatim Reports</td>
<td>9.461</td>
<td>0.284</td>
<td>0.943</td>
</tr>
<tr>
<td>Activity Changes WITH Secondary Activities</td>
<td>1.500</td>
<td>0.126</td>
<td>0.393</td>
</tr>
<tr>
<td><strong>Intercepts</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of Entries</td>
<td>12.197</td>
<td>0.308</td>
<td>1.906</td>
</tr>
<tr>
<td>Total number of Secondary Activities</td>
<td>0.841</td>
<td>0.054</td>
<td>0.541</td>
</tr>
<tr>
<td>Total number of Activity Changes</td>
<td>3.133</td>
<td>0.134</td>
<td>0.998</td>
</tr>
<tr>
<td>Total number of Where Changes</td>
<td>1.894</td>
<td>0.228</td>
<td>0.563</td>
</tr>
<tr>
<td>Total number of Verbatim Reports</td>
<td>13.954</td>
<td>0.404</td>
<td>1.391</td>
</tr>
<tr>
<td><strong>Variance</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rapport</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Residual Variances</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of Entries</td>
<td>3.645</td>
<td>0.325</td>
<td>0.089</td>
</tr>
<tr>
<td>Total number of Secondary Activities</td>
<td>2.136</td>
<td>0.249</td>
<td>0.883</td>
</tr>
<tr>
<td>Total number of Activity Changes</td>
<td>6.813</td>
<td>0.276</td>
<td>0.69</td>
</tr>
<tr>
<td>Total number of Where Changes</td>
<td>9.713</td>
<td>1.475</td>
<td>0.857</td>
</tr>
<tr>
<td>Total number of Verbatim Reports</td>
<td>11.151</td>
<td>1.733</td>
<td>0.111</td>
</tr>
<tr>
<td><strong>R-Square</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of Entries</td>
<td>0.911</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of Secondary Activities</td>
<td>0.117</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of Activity Changes</td>
<td>0.31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of Where Changes</td>
<td>0.143</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of Verbatim Reports</td>
<td>0.889</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Two-step approach for SEM.** Before proceeding to estimate the structural model where data quality is predicted by survey, a confirmatory factor analysis with the three factors (two for data quality and one for interview rapport) was estimated. The reliability and dimensionality of the 11 items was assessed in the same sample of 12762 respondents using Mean and Variance adjusted Weighted Least Square (WLSMV) estimator and Theta scaling in Mplus v. 7.11. All models were identified by setting the latent factor mean to 0 and the latent factor variance to 1, such that all thresholds and item factor loadings estimated. Additionally, the CLUSTER option under TYPE=COMPLEX was used to indicate that respondents are nested within 69 interviewers in this dataset. The use of these options in Mplus did not permit to define the variables with a Poisson distribution as count but they needed to be treated as continuous.

Table 6.7 includes the obtained model $\chi^2$, its degrees of freedom, and its $p$-value, CFI, and the RMSEA point estimate and 90% confidence interval. Nested model comparisons, shown in Table 6.8, were conducted the $\chi^2$ Test for Difference Testing implemented by Mplus using the DIFFTEST option. As previously stated, given the large number of observations, the cut-off level of significance for these comparisons is $p<.01$. The specific models examined are described in detail below.

The three factor model with 11 indicators was not an exact representation of the data, as $\chi^2 (41) = 1524.002, p<.05$. Likewise, the model did not have an approximate or close fit to the data as RMSEA=0.053, $p=.010$, and CFI=.845. Thus, this model is not adequate according to any of the fit criteria. Sources of local misfit were identified using
the normalized residual covariance matrix that indicated a large negative residual covariance between activity changes and secondary activity reports (a covariance already added in the interview rapport measurement model). When this parameter was added, the model failed to estimate. Looking at residual variances, it was diagnosed that number of entries was the indicator causing the estimation failure. The reason for this being that ultimately, number of entries are a function of the number of activities reported, and thus, when these two variables are simultaneously used in a model, their linear dependency results in a non-positive definite residual covariance matrix.

Number of entries was thus removed from the model, and a three-factor model was estimated anew. Still the model did not exactly fit the data, as $\chi^2 (32) = 1160.515$, nor had an approximate fit to the data, as CFI = .935. The model did, however, showed to have a closer fit to the data as the RMSEA=.053, $p=.050$. Sources of local misfit again indicated a large negative residual covariance between activity changes and secondary activity reports, so this a covariance that had already been present in the interview rapport CFA model, was added to the model. The fit of the model increased somewhat; although the model did not exactly fit the data, as $\chi^2 (31) = 520.403$, the close and approximate fit indices improved, as CFI = .936 and RMSEA=.035, $p=1.0$. Just as it occurred in the data quality measurement model, sources of local misfit again indicated that Missing a basic activity and Number of reported activities were correlated. When this parameter was added, the model seemed to adequately fit the data according to the approximate and close fit indices, for CFI=.954 and RMSEA=.030, $p=1.00$. The model did not fit the data exactly still ($\chi^2 (31) = 378.773$), but this was expected due to the large sample size. An
additional covariance suggested my MODINDICES was the one between *missing who* and *missing where*; a model with that added covariance, however, was not estimable. Further examination of local fit via normalized residual covariances and modification indices yielded no remaining relations that were theoretically justifiable. Importantly, given that the model had achieved a good fit, I decided not to allow any cross-loadings in order to be able to measure data quality and interview rapport separately.

Table 6.9 provides estimates and their standard errors for the item factor loadings and thresholds, from both the unstandardized and standardized solutions, together with the R-squares for the observed variables. All factor loadings and the factor covariance were statistically significant. As shown in Table 6.9, standardized loadings ranged from .302 to .934 (with $R^2$ values for the amount of item variance accounted for by the factor ranging from .091 to .871), suggesting the factor loadings were practically significant as well.
Table 6.7
Assessment of Model Fit using WLSMV for Three-Factor Measurement Model

<table>
<thead>
<tr>
<th>Model</th>
<th># Items</th>
<th>$\chi^2$</th>
<th>$\chi^2$ DF</th>
<th>$\chi^2$ p-value</th>
<th>CFI</th>
<th>RMSEA Lower CI</th>
<th>RMSEA Higher CI</th>
<th>RMSEA p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Three-Factors</td>
<td>11</td>
<td>1524.002</td>
<td>41</td>
<td>&lt;.0001</td>
<td>.845</td>
<td>.053</td>
<td>.051</td>
<td>.056</td>
</tr>
<tr>
<td>Three-Factors (Activity changes with secondary activities)</td>
<td>11</td>
<td>Did not estimate</td>
<td>31</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Three-Factors (Removed entries)</td>
<td>10</td>
<td>1160.515</td>
<td>32</td>
<td>&lt;.0001</td>
<td>.852</td>
<td>.053</td>
<td>.050</td>
<td>.055</td>
</tr>
<tr>
<td>Three-Factors (Activity changes with secondary activities, entries removed)</td>
<td>10</td>
<td>520.404</td>
<td>31</td>
<td>&lt;.0001</td>
<td>.936</td>
<td>.035</td>
<td>.033</td>
<td>.038</td>
</tr>
<tr>
<td>Three-Factor (Missing basic activities with number of activities)</td>
<td>10</td>
<td>377.850</td>
<td>30</td>
<td>&lt;.0001</td>
<td>.954</td>
<td>.030</td>
<td>.027</td>
<td>.033</td>
</tr>
</tbody>
</table>

*Note. WLSMV=Weighted Least Squares; CFI=Comparative Fit Index; RMSEA=Root mean square error of approximation.*

Table 6.8
Nested Model Comparisons for Three-Factor Measurement Model

<table>
<thead>
<tr>
<th>Model</th>
<th>DF difference</th>
<th>$\chi^2$ Test Difference</th>
<th>$\chi^2$ Test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Three-Factors (Removed entries)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Three-Factors (Activity changes with secondary activities, entries removed)</td>
<td>1</td>
<td>1491.650</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Three-Factor (Missing basic activities with number of activities)</td>
<td>1</td>
<td>372.834</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

*Note. DF=Degrees of Freedom.*
Table 6.9

*Weighted Least Squares Estimates of Factor Loadings for a Three-Factor Measurement Model*

<table>
<thead>
<tr>
<th>Model Parameters</th>
<th>Unstandardized Estimate</th>
<th>SE</th>
<th>Standardized Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MOTIVATION BY</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of activities</td>
<td>6.345</td>
<td>0.131</td>
<td>0.812</td>
</tr>
<tr>
<td>Not missing basic activity reports</td>
<td>0.317</td>
<td>0.021</td>
<td>0.302</td>
</tr>
<tr>
<td>Duration of the interview in minutes</td>
<td>5.293</td>
<td>0.134</td>
<td>0.905</td>
</tr>
<tr>
<td><strong>MEMORY BY</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Missing who code</td>
<td>0.576</td>
<td>0.026</td>
<td>0.499</td>
</tr>
<tr>
<td>No Missing where code</td>
<td>0.590</td>
<td>0.031</td>
<td>0.508</td>
</tr>
<tr>
<td>No ATUS error code</td>
<td>0.413</td>
<td>0.034</td>
<td>0.382</td>
</tr>
<tr>
<td><strong>RAPPORT BY</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of Secondary Activities</td>
<td>0.520</td>
<td>0.017</td>
<td>0.334</td>
</tr>
<tr>
<td>Total number of Activity Changes</td>
<td>1.845</td>
<td>0.057</td>
<td>0.587</td>
</tr>
<tr>
<td>Total number of Where Changes</td>
<td>1.241</td>
<td>0.065</td>
<td>0.369</td>
</tr>
<tr>
<td>Total number of Verbatim Reports</td>
<td>9.366</td>
<td>0.141</td>
<td>0.934</td>
</tr>
<tr>
<td><strong>MOTIVATION WITH</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEMORY</td>
<td>-0.713</td>
<td>0.036</td>
<td>-0.713</td>
</tr>
<tr>
<td><strong>RAPPORT WITH</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOTIVATION</td>
<td>0.872</td>
<td>0.012</td>
<td>0.872</td>
</tr>
<tr>
<td>MEMORY</td>
<td>-0.747</td>
<td>0.027</td>
<td>-0.747</td>
</tr>
<tr>
<td>Interview duration with not missing basic activities</td>
<td>1.817</td>
<td>0.083</td>
<td>0.399</td>
</tr>
<tr>
<td>Activity changes with Number of Secondary Activities</td>
<td>1.472</td>
<td>0.027</td>
<td>0.395</td>
</tr>
<tr>
<td><strong>Intercepts</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of activities</td>
<td>19.419</td>
<td>0.299</td>
<td>2.486</td>
</tr>
<tr>
<td>Interview duration in minutes</td>
<td>10.630</td>
<td>0.439</td>
<td>1.817</td>
</tr>
<tr>
<td>Total number of Secondary Activities</td>
<td>0.841</td>
<td>0.081</td>
<td>0.541</td>
</tr>
<tr>
<td>Total number of Activity Changes</td>
<td>3.132</td>
<td>0.178</td>
<td>0.997</td>
</tr>
<tr>
<td>Total number of Where Changes</td>
<td>1.894</td>
<td>0.303</td>
<td>0.563</td>
</tr>
<tr>
<td>Total number of Verbatim Reports</td>
<td>13.954</td>
<td>0.465</td>
<td>1.391</td>
</tr>
</tbody>
</table>
Table 6.9 (continued)

<table>
<thead>
<tr>
<th>Thresholds</th>
<th>Factor Loadings</th>
<th>Standard Errors</th>
<th>Variances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Missing Basic Activities$1</td>
<td>-0.727</td>
<td>0.036</td>
<td>-0.693</td>
</tr>
<tr>
<td>Not Who Missing Codes $1</td>
<td>-0.410</td>
<td>0.062</td>
<td>-0.356</td>
</tr>
<tr>
<td>Not Where Missing Codes $1</td>
<td>-0.005</td>
<td>0.063</td>
<td>-0.004</td>
</tr>
<tr>
<td>Not ATUS Error Codes$1</td>
<td>-0.955</td>
<td>0.073</td>
<td>-0.883</td>
</tr>
</tbody>
</table>

Variance

<table>
<thead>
<tr>
<th>Observed Variable</th>
<th>Motivation</th>
<th>Memory</th>
<th>Rapport</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of activities</td>
<td>1.000</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Interview duration in minutes</td>
<td>6.211</td>
<td>0.798</td>
<td>0.181</td>
</tr>
<tr>
<td>Total number of Secondary Activities</td>
<td>2.150</td>
<td>0.016</td>
<td>0.888</td>
</tr>
<tr>
<td>Total number of Activity Changes</td>
<td>6.462</td>
<td>0.150</td>
<td>0.655</td>
</tr>
<tr>
<td>Total number of Where Changes</td>
<td>9.793</td>
<td>0.497</td>
<td>0.864</td>
</tr>
<tr>
<td>Total number of Verbatim Reports</td>
<td>12.938</td>
<td>1.377</td>
<td>0.129</td>
</tr>
</tbody>
</table>

R-SQUARE

<table>
<thead>
<tr>
<th>Observed Variable</th>
<th>Total number of activities</th>
<th>Not Missing Basic Activities</th>
<th>Who Missing Codes</th>
<th>Where Missing Codes</th>
<th>Interview duration in minutes</th>
<th>ATUS Error Codes</th>
<th>Total number of Secondary Activities</th>
<th>Total number of Activity Changes</th>
<th>Total number of Where Changes</th>
<th>Total number of Verbatim Reports</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of activities</td>
<td>0.660</td>
<td>0.091</td>
<td>0.249</td>
<td>0.258</td>
<td>0.819</td>
<td>0.146</td>
<td>0.112</td>
<td>0.345</td>
<td>0.136</td>
<td>0.871</td>
</tr>
<tr>
<td>Not Missing Basic Activities</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Who Missing Codes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Where Missing Codes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. SE=Standard error.

**Predicting data quality from interview rapport.** Once the measurement model was estimated and it fit the data well, a full structural regression model was estimated using MLSEMV estimation in Mplus 7.11, to predicted motivation to respond and memory availability (as the two data quality factors) from interview rapport. In this case, all
models were identified by the marker indicator technique (setting the loading of one of the indicators to 1). All factor loadings were statistically significant, as well as the factor covariance between motivation and memory. The fit of the model was good as indicated by the approximate and close fit indices (CFI=0.954 and RMSEA=0.030, $p=1.000$). The model did not fit the data exactly, as $\chi^2 (31) = 378.773$, but this was expected due to the large sample size.

Table 16.10 provides estimates and their standard errors for the item factor loadings and thresholds, from both the unstandardized and standardized solutions, as well as the R-squares for observed and latent variables. As shown in Table 6.10, standardized loadings ranged from .299 to .944 (with $R^2$ values for the amount of item variance accounted for by the factor ranging from .089 to .892), suggesting the factor loadings were practically significant as well.

As can be seen, interview rapport significantly predicts both quality factors, though in opposite directions. The first data quality factor, related to total number of activities, not having failed to report at least one basic activity, and the duration of the interview in minutes is positively and significantly predicted by interview rapport. The second data quality, associated with having failed to provide information on details such as who was present or where the activity took place, as well as not having any of the ATUS error codes, is negatively predicted by interview rapport. Thus, whereas interview rapport appears beneficial to the motivation factor, it seems to be detrimental to the quality factor thought to be governed by memory processes. A possible explanation for such counterintuitive results is discussed in the final section of this dissertation. Figure
6.14 shows the diagram of the estimated structural equation model predicting data quality from interview rapport.

Table 6.10
*Weighted Least Squares Estimates of Factor Loadings for a Structural Equation Model predicting Motivation and Memory from Interview Rapport*

<table>
<thead>
<tr>
<th>Model Parameters</th>
<th>Unstandardized Estimate</th>
<th>SE</th>
<th>Standardized Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MOTIVATION BY</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of activities</td>
<td>1.000</td>
<td>0.000</td>
<td>0.796</td>
</tr>
<tr>
<td>Not missing basic activity reports</td>
<td>0.050</td>
<td>0.003</td>
<td>0.299</td>
</tr>
<tr>
<td>Duration of the interview in minutes</td>
<td>0.842</td>
<td>0.025</td>
<td>0.903</td>
</tr>
<tr>
<td><strong>MEMORY BY</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Missing <em>who</em> code</td>
<td>1.000</td>
<td>0.000</td>
<td>0.499</td>
</tr>
<tr>
<td>No Missing <em>where</em> code</td>
<td>1.023</td>
<td>0.049</td>
<td>0.507</td>
</tr>
<tr>
<td>No ATUS error code</td>
<td>0.717</td>
<td>0.056</td>
<td>0.382</td>
</tr>
<tr>
<td><strong>RAPPORT BY</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of Verbatim Reports</td>
<td>1.000</td>
<td>0.000</td>
<td>0.944</td>
</tr>
<tr>
<td>Total number of Secondary Activities</td>
<td>0.055</td>
<td>0.002</td>
<td>0.333</td>
</tr>
<tr>
<td>Total number of Activity Changes</td>
<td>0.194</td>
<td>0.007</td>
<td>0.586</td>
</tr>
<tr>
<td>Total number of Where Changes</td>
<td>0.131</td>
<td>0.008</td>
<td>0.368</td>
</tr>
<tr>
<td><strong>MOTIVATION ON RAPPORT</strong></td>
<td>0.582</td>
<td>0.013</td>
<td>0.881</td>
</tr>
<tr>
<td><strong>MEMORY ON RAPPORT</strong></td>
<td>-0.045</td>
<td>0.002</td>
<td>-0.741</td>
</tr>
<tr>
<td><strong>MOTIVATION WITH MEMORY</strong></td>
<td>-0.245</td>
<td>0.100</td>
<td>-0.214</td>
</tr>
<tr>
<td>Activity Changes with Number of Secondary Activities</td>
<td>1.478</td>
<td>0.027</td>
<td>0.396</td>
</tr>
<tr>
<td>Not Missing basic activities with Number of activities reported</td>
<td>1.864</td>
<td>0.082</td>
<td>0.391</td>
</tr>
<tr>
<td><strong>Intercepts</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of activities</td>
<td>19.416</td>
<td>0.299</td>
<td>2.468</td>
</tr>
<tr>
<td>Interview duration in minutes</td>
<td>10.630</td>
<td>0.439</td>
<td>1.823</td>
</tr>
</tbody>
</table>
### Table 6.10 (continued)

<table>
<thead>
<tr>
<th>Model Parameters</th>
<th>Unstandardized Estimate</th>
<th>SE</th>
<th>Standardized Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of Secondary Activities</td>
<td>0.841</td>
<td>0.081</td>
<td>0.541</td>
</tr>
<tr>
<td>Total number of Activity Changes</td>
<td>3.133</td>
<td>0.178</td>
<td>0.998</td>
</tr>
<tr>
<td>Total number of Where Changes</td>
<td>1.894</td>
<td>0.303</td>
<td>0.563</td>
</tr>
<tr>
<td>Total number of Verbatim Reports</td>
<td>13.954</td>
<td>0.465</td>
<td>1.392</td>
</tr>
</tbody>
</table>

**Thresholds**

| Not Missing Basic Activities$1               | -0.727                  | 0.036 | -0.694               |
| No Missing Who Codes $1                      | -0.409                  | 0.062 | -0.355               |
| No Missing Where Codes $1                    | -0.005                  | 0.063 | -0.004               |
| Not ATUS Error Codes$1                       | -0.956                  | 0.073 | -0.884               |

**Variances**

| Rapport                                      | 89.665                  | 2.674 | 1                    |

**Residual Variances**

| Total number of activities                   | 22.710                  | 0.906 | 0.367                |
| Interview duration in minutes                | 6.254                   | 0.794 | 0.184                |
| Total number of Secondary Activities         | 2.151                   | 0.016 | 0.889                |
| Total number of Activity Changes             | 6.480                   | 0.151 | 0.657                |
| Total number of Where Changes                | 9.788                   | 0.497 | 0.865                |
| Total number of Verbatim Reports             | 10.851                  | 1.410 | 0.108                |
| Quality 1                                    | 8.801                   | 0.919 | 0.225                |
| Quality 2                                    | 0.149                   | 0.023 | 0.451                |

**R-SQUARE**

| Observed Variable                            |                         |      |                      |
| Total number of activities                   | 0.633                   | 0.015 |                      |
| Not Missing Basic Activities                 | 0.089                   | 0.011 |                      |
| Who Missing Codes                            | 0.249                   | 0.017 |                      |
| Where Missing Codes                          | 0.257                   | 0.020 |                      |
| Interview duration in minutes                | 0.816                   | 0.023 |                      |
| ATUS Error Codes                             | 0.146                   | 0.020 |                      |
| Total number of Secondary Activities         | 0.111                   | 0.007 |                      |
| Total number of Activity Changes             | 0.343                   | 0.014 |                      |
| Total number of Where Changes                | 0.135                   | 0.015 |                      |
| Total number of Verbatim Reports             | 0.892                   | 0.014 |                      |

| Latent Variable                              |                         |      |                      |
| Motivation                                   | 0.775                   | 0.021 |                      |
| Memory                                       | 0.549                   | 0.040 |                      |
Figure 6.14. Diagram of the Estimated Structural Equation Model Predicting Data Quality from Interview Rapport
Inclusion of control variables in the prediction of Motivation, Memory, and Interview Rapport. Once the full structural equation model was estimated predicting data quality from interview rapport, an analysis was conducted where control variables hypothesized to impact both factors of data quality and that of interview rapport were included. Table 6.11 includes the estimates for the analysis done in Mplus v. 7.11 in which all three factors were simultaneously predicted from respondent’s demographics, timing of the interview, interviewer’s cooperation rate, respondent’s busyness, social capital, and participation reluctance.

As can be seen from Table 6.11, whereas age positively predicts motivation to respond in a thorough manner, age negatively predicts having the memory resources to answer with enough precision. If we take into account that the range of age in the 2010 ATUS is 70 years (as persons ranging from 15-year olds to about 85-year olds responded to the survey), this effect may be important to take into consideration. It seems that compared to younger respondents, older respondents are more willing to optimize and make an effort to provide a more complete report of their activities while spending more time on the interview. At the same time, it was found that, compared to younger people, older people may also face more cognitive challenges and fail to provide full details such as with whom they were and where they were during a particular activity. Age was not found to be a significant predictor of rapport. Gender has a significant effect on the three latent factors. Men are expected to have a lower degree of motivation and rapport than women, indicating that women were more likely to place more effort in building rapport and reporting on activities from the previous day in comparison to men. Men, however,
seem to fare better when it comes to remembering details of their activities than women. All factors are affected by education. Compared to those respondent without a high school diploma, respondents with some college education, a bachelor’s degree, or a graduate degree seemed to be more motivated and to have more memory issues. Likewise, compared to respondents without high school education, it seems that more rapport is built with respondents who are more educated.

The number of hours elapsed before the interview took place was not a significant predictor of motivation or memory. It did, however, significantly predict rapport: The more hours elapsed, the lower the degree of rapport. The cooperation rate of interviewers significantly predicted motivation and memory. The higher the cooperation rate, the higher the motivation, but the lower the memory. This result is consistent with previous research that argues that the more experienced interviewers are, in fact, able to better motivate respondents, but as they get more experienced (or fatigued with their job), the more likely they are to stop probing for additional details and to accelerate the pace with which the conduct the interviews (Olson and Peytchev, 2007; Olson and Bilgen, 2011).

As for the busyness variables, they were significantly predictors of all three factors. As expected, compared to people who are not in the labor force, motivation and rapport for respondents in any occupation was lower, whereas having an executive type of occupation was found positively related to having memory resources to give additional details of the reported activities. Interestingly, the higher the percent of adults who work in the household, the higher the motivation and the rapport.
With regard to the social capital variables, as predicted by the theory, all three were significant positive predictors of motivation; among respondents living in households with a higher number of members, where young children (under the age of 6) were present, and where the respondent was married, it was expected to find a higher degree of motivation to provide complete answers. Likewise, respondents with a higher number of household members and where there are young children present, the degree of rapport is higher. On the contrary, respondents with a higher number of family members and who have small children in the household seemed to have less memory resources to provide additional details about their activities, compared to those with a lower number of family members and where there are no young children present.
Table 6.11

Regression Estimates of the Prediction of Motivation, Memory, and Rapport

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Motivation</th>
<th></th>
<th>Memory</th>
<th></th>
<th>Interview Rapport</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>SE</td>
<td>Estimate</td>
<td>SE</td>
<td>Estimate</td>
<td>SE</td>
</tr>
<tr>
<td>Demographics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Age (Centered at 47 years old)</td>
<td>0.014</td>
<td>0.005</td>
<td>-0.002</td>
<td>0.001</td>
<td>-0.060</td>
<td>0.016</td>
</tr>
<tr>
<td>- Gender (Reference: Male)</td>
<td>-2.604</td>
<td>0.161</td>
<td>0.133</td>
<td>0.022</td>
<td>0.120</td>
<td>-2.699</td>
</tr>
<tr>
<td>Education (Ref: Less than high school)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- High school education</td>
<td>0.307</td>
<td>0.232</td>
<td>-0.027</td>
<td>0.032</td>
<td>-0.022</td>
<td>1.099</td>
</tr>
<tr>
<td>- Some college education</td>
<td>1.232</td>
<td>0.240</td>
<td>0.083</td>
<td>0.033</td>
<td>-0.074</td>
<td>2.778</td>
</tr>
<tr>
<td>- Bachelor degree</td>
<td>1.922</td>
<td>0.220</td>
<td>0.116</td>
<td>0.036</td>
<td>-0.084</td>
<td>4.037</td>
</tr>
<tr>
<td>- Graduate degree</td>
<td>2.698</td>
<td>0.228</td>
<td>0.131</td>
<td>0.043</td>
<td>-0.099</td>
<td>5.175</td>
</tr>
<tr>
<td>Timing of the interview</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Hours elapsed before the interview</td>
<td>0.043</td>
<td>0.042</td>
<td>0.026</td>
<td>0.010</td>
<td>0.013</td>
<td>-0.121</td>
</tr>
<tr>
<td>Interviewer's cooperation rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Cooperation Rate in the 2010 ATUS</td>
<td>6.302</td>
<td>1.850</td>
<td>0.086</td>
<td>0.392</td>
<td>-0.234</td>
<td>5.960</td>
</tr>
<tr>
<td>Busyness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Percent of household adults who work</td>
<td>0.686</td>
<td>0.193</td>
<td>0.042</td>
<td>0.026</td>
<td>-0.026</td>
<td>0.892</td>
</tr>
</tbody>
</table>
Table 6.11 (continued)

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Motivation Estimate</th>
<th>Motivation Standard Estimate</th>
<th>Memory Estimate</th>
<th>Memory Standard Estimate</th>
<th>Interview Rapport Estimate</th>
<th>Interview Rapport Standard Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of occupation (Ref.: Not in the labor force)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Executive/Professional</td>
<td>-0.996</td>
<td>0.240</td>
<td>0.088</td>
<td>0.032</td>
<td>-1.325</td>
<td>0.343</td>
</tr>
<tr>
<td></td>
<td>-0.524</td>
<td>0.217</td>
<td>0.038</td>
<td>0.037</td>
<td>-1.176</td>
<td>0.408</td>
</tr>
<tr>
<td>- Support/Production</td>
<td>-0.551</td>
<td>0.186</td>
<td>0.057</td>
<td>0.031</td>
<td>-0.927</td>
<td>0.286</td>
</tr>
<tr>
<td>Social capital</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Household size</td>
<td>0.515</td>
<td>0.055</td>
<td>-0.058</td>
<td>0.010</td>
<td>0.545</td>
<td>0.093</td>
</tr>
<tr>
<td></td>
<td>0.532</td>
<td>0.141</td>
<td>-0.031</td>
<td>0.018</td>
<td>-0.028</td>
<td>0.037</td>
</tr>
<tr>
<td></td>
<td>1.515</td>
<td>0.127</td>
<td>-0.088</td>
<td>0.021</td>
<td>1.498</td>
<td>0.202</td>
</tr>
<tr>
<td>Participation Reluctance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Number of weeks before interview was completed</td>
<td>-0.198</td>
<td>0.034</td>
<td>0.008</td>
<td>0.005</td>
<td>0.023</td>
<td>0.054</td>
</tr>
</tbody>
</table>

Note. Significant estimates ($p<0.05$) are in bold.
Fidelity Analysis

In this analysis a multiple regression model was estimated predicting overall health taken as a proxy of overall wellbeing, but where factor scores obtained from the previously conducted measurement models were included as weights. The way these factor scores were constructed and their transformation in survey weights was explained in Chapter 5. As previously explained, following Juster et al. (1985) and Krueger et al. (2009), the independent variables used to predict overall health included household income, living arrangements (household is rented or owned), marital status, employment status, number of children, activity-based wellbeing measures taken from the answers to the questions in the wellbeing module where respondents were asked how they felt the day before doing specific activities (activity was meaningful, happy, sad, stressful, tiresome, painful). A second set of models was estimated controlling for the influence of education.

The SURVEY REG procedure with the CLUSTER and WEIGHT options in SAS 9.4 was used. As deemed necessary because of the level of interviewer variance (ICC for overall health was 0.01, with a resulting average design effect of 2.7), the CLUSTER option was used to account for the influence of interviewers, whereas the WEIGHT option was used to incorporate the effect of factor scores of the two types of data quality (motivation and memory) and interview rapport.

Table 6.12 includes the estimates, standard errors and R-squares for four different models predicting overall health. The first model predicts overall health by using the CLUSTER option, but without using any factor score weights, and the three next models
predict overall health using the CLUSTER option and each of the factor scores used as weights in this order: motivation factor scores, memory factor scores, and rapport factor scores. Table 16.13 includes the estimates, standard errors and R-squares for four different models, this time controlling for respondent’s education.

As can be seen from Table 16.12, compared to the other models, the model in which the memory factor scores were used as weights has the highest R-square (.276 compared to .269, .255, and .253 for the model without weights, the one using motivation weights, and the one using rapport weights, respectively). The same pattern is displayed in the models where education is included in the model. Thus, compared to a model without weights, the model that uses the memory weights has a 2.6% increase in the total amount of explained variance. Likewise, compared to the model using the motivation weights, the memory weights model has an increase of 8.2% in the total amount of explained variance of overall health. Finally, compared to the rapport weights model, the memory weights model has an increase of 9.1% of the amount of overall health variance explained.

A final point to take into consideration in this analysis is that in the first set of four models estimated, the only difference regarding the significance of estimates across models is that in the memory weights model, the number of children living in the household is a significant predictor of overall health (in bold in table 16.12). In this case, for an additional child in the family, the overall health score is predicted to decrease by 0.020 units. However, as Table 16.13 shows, once education was added as a covariate, such difference regarding the covariate of number of children is not present any more. In
effect, the estimates and the level of significance for each of the variables are practically
the same across all four models to the first decimal. Thus, it appears that when
controlling for education, any influence of the quality of the data is removed, at least
when predicting overall health.
Table 16.12

*Regression Model Comparing Models using Different Factors as Weights*

<table>
<thead>
<tr>
<th>Variable</th>
<th>No weights used</th>
<th>Motivational Factor</th>
<th>Memory Factor</th>
<th>Rapport Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standardized Estimate</td>
<td>SE</td>
<td>Standardized Estimate</td>
<td>SE</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.000</td>
<td>0.044</td>
<td>0.000</td>
<td>0.050</td>
</tr>
<tr>
<td>Family Income</td>
<td>0.167</td>
<td>0.000</td>
<td>0.175</td>
<td>0.000</td>
</tr>
<tr>
<td>Marital Status (Not Married=0)</td>
<td>-0.001</td>
<td>0.026</td>
<td>-0.007</td>
<td>0.026</td>
</tr>
<tr>
<td>Employment Status (Unemployed=0)</td>
<td>0.112</td>
<td>0.020</td>
<td>0.100</td>
<td>0.022</td>
</tr>
<tr>
<td>Number of Children</td>
<td>-0.015</td>
<td>0.007</td>
<td>-0.003</td>
<td>0.008</td>
</tr>
<tr>
<td>Average of 'Happy' Activities</td>
<td>0.093</td>
<td>0.010</td>
<td>0.080</td>
<td>0.009</td>
</tr>
<tr>
<td>Average of 'Meaningful' Activities</td>
<td>0.003</td>
<td>0.007</td>
<td>0.002</td>
<td>0.008</td>
</tr>
<tr>
<td>Average of 'Painful' Activities</td>
<td>-0.263</td>
<td>0.006</td>
<td>-0.266</td>
<td>0.008</td>
</tr>
<tr>
<td>Average of 'Sad' Activities</td>
<td>-0.021</td>
<td>0.012</td>
<td>-0.026</td>
<td>0.012</td>
</tr>
<tr>
<td>Average of 'Stressed' Activities</td>
<td>-0.041</td>
<td>0.008</td>
<td>-0.034</td>
<td>0.008</td>
</tr>
<tr>
<td>Average of 'Tiresome' Activities</td>
<td>-0.101</td>
<td>0.007</td>
<td>-0.099</td>
<td>0.008</td>
</tr>
<tr>
<td>Living arrangements (Rented=0)</td>
<td>0.046</td>
<td>0.025</td>
<td>0.047</td>
<td>0.027</td>
</tr>
<tr>
<td>Gender (Female=0)</td>
<td>-0.031</td>
<td>0.015</td>
<td>-0.034</td>
<td>0.016</td>
</tr>
<tr>
<td>Age (Centered at mean 47)</td>
<td>-0.147</td>
<td>0.001</td>
<td>-0.130</td>
<td>0.001</td>
</tr>
<tr>
<td>R-square</td>
<td>.269</td>
<td>.255</td>
<td>.276</td>
<td>.253</td>
</tr>
</tbody>
</table>

Note: SE= Standard Error. Significant differences in estimates across models are in bold.
Table 16.13

Regression model comparing models using different factors as weights controlling for education

<table>
<thead>
<tr>
<th>Variable</th>
<th>No weights used</th>
<th>Motivational Factor</th>
<th>Memory Factor</th>
<th>Rapport Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standardized</td>
<td>Standardized</td>
<td>Standardized</td>
<td>Standardized</td>
</tr>
<tr>
<td></td>
<td>Estimate</td>
<td>SE</td>
<td>Estimate</td>
<td>SE</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.000</td>
<td>0.050</td>
<td>0.000</td>
<td>0.054</td>
</tr>
<tr>
<td>Family Income</td>
<td>0.116</td>
<td>0.000</td>
<td>0.120</td>
<td>0.000</td>
</tr>
<tr>
<td>Marital Status (Not Married=0)</td>
<td>-0.011</td>
<td>0.026</td>
<td>-0.020</td>
<td>0.026</td>
</tr>
<tr>
<td>Employment Status (Unemployed=0)</td>
<td>0.086</td>
<td>0.019</td>
<td>0.074</td>
<td>0.022</td>
</tr>
<tr>
<td>Number of Children</td>
<td>-0.005</td>
<td>0.008</td>
<td>0.007</td>
<td>0.008</td>
</tr>
<tr>
<td>Average of 'Happy' Activities</td>
<td>0.099</td>
<td>0.009</td>
<td>0.089</td>
<td>0.009</td>
</tr>
<tr>
<td>Average of 'Meaningful' Activities</td>
<td>0.007</td>
<td>0.007</td>
<td>0.004</td>
<td>0.008</td>
</tr>
<tr>
<td>Average of 'Painful' Activities</td>
<td>-0.253</td>
<td>0.006</td>
<td>-0.257</td>
<td>0.008</td>
</tr>
<tr>
<td>Average of 'Sad' Activities</td>
<td>-0.011</td>
<td>0.012</td>
<td>-0.014</td>
<td>0.012</td>
</tr>
<tr>
<td>Average of 'Stressed' Activities</td>
<td>-0.052</td>
<td>0.008</td>
<td>-0.046</td>
<td>0.008</td>
</tr>
<tr>
<td>Average of 'Tiresome' Activities</td>
<td>-0.099</td>
<td>0.007</td>
<td>-0.097</td>
<td>0.007</td>
</tr>
<tr>
<td>Living arrangements (Rented=0)</td>
<td>0.052</td>
<td>0.024</td>
<td>0.052</td>
<td>0.026</td>
</tr>
<tr>
<td>Gender (Female=0)</td>
<td>-0.023</td>
<td>0.014</td>
<td>-0.025</td>
<td>0.016</td>
</tr>
<tr>
<td>Age (Centered at mean 47)</td>
<td>-0.165</td>
<td>0.001</td>
<td>-0.151</td>
<td>0.001</td>
</tr>
<tr>
<td>High school</td>
<td>0.035</td>
<td>0.035</td>
<td>0.040</td>
<td>0.037</td>
</tr>
<tr>
<td>Some college</td>
<td>0.083</td>
<td>0.040</td>
<td>0.091</td>
<td>0.040</td>
</tr>
<tr>
<td>Bachelor degree</td>
<td>0.141</td>
<td>0.042</td>
<td>0.156</td>
<td>0.044</td>
</tr>
<tr>
<td>Graduate degree</td>
<td>0.129</td>
<td>0.038</td>
<td>0.142</td>
<td>0.038</td>
</tr>
<tr>
<td>R-square</td>
<td>0.284</td>
<td>0.272</td>
<td>0.29</td>
<td>0.269</td>
</tr>
</tbody>
</table>

Note: SE= Standard Error. Significant differences in estimates across models are in bold.
Summary of Findings

Indexes of ‘Data Quality’ and Rapport. Contrary to what was initially hypothesized, two data quality factors, rather than one, were identified. The first factor appears to be related to the respondent’s willingness or motivation to provide thorough answers, whereas the second one seems to be governed by memory processes. The factor loadings for the indicators of both factors were all statistically significant, though variance explained by the factor was higher for the indicators of the motivation factor. The time taken to respond and the number of activities reported are the most discriminant indicators of motivation. In the memory factor, all three indicators seem to be equally discriminant, although this assertion still needs to be tested. Further, it should be noted that both underlying processes (motivation and memory) may occur at the same time; in effect, a respondent may be willing to provide a thorough report, but because of cognitive challenges he or she simply cannot remember the additional details of an activity that occurred during the previous day, and thus fail to provide a precise report.

With regard to the construct of interview rapport, a one-factor CFA model seemed to fit the data well. Rapport was indicated by the total number of entries (which do not necessarily coincide with the number of activities reported), number of activity and where changes, number of secondary activities reported, and number of verbatim reports. All the factor loadings of the one-factor CFA model were statistically and practically significant, being the number of entries and the number of verbatim reports the most
discriminant indicators of this construct. The conceptual significance of the indicator of the number of secondary activities reported should be highlighted, as the ATUS does not require interviewers to probe about secondary activities, as this information is never released in the public datasets that the Bureau of Labor Statistics produces. However, the fact that these activities were recorded are taken as a signal that a much richer conversation took place, one that was enabled by a higher level of rapport.

In any event, the present work, exploratory in nature, seems to have found additional evidence to the affirmation that quality is a multidimensional concept, which should be studied in light of the different survey errors put forward in the literature. Indeed, within that framework, the fact that two different quality factors were identified is not surprising.

A brief discussion of reflective vs formative models. It seems to be the case that the factors that were identified reflect the behaviors and capacities of respondents, such as their motivation to provide complete answers and the memory resources they have in order to provide additional details to the reported activities (who with and where). What is more, following previous research (Fricker, 20007), I believe that, the total number of activity reports or the indicator of not having failed to report basic activities (e.g., sleeping or eating), are a reflection of the first data quality factor, the respondent’s willingness to report, and not necessarily the actual reality of what happened during the respondent’s day. In other words, a respondent with many activities to report, may simply decide to satisfice (Krosnick, 1991) and fail to report all of the activities that took place, whereas a motivated respondent will decide to provide more information and take more
time to do so. With the second data quality factor, the “quality” that is measured is reflected by a lower number of failures to report an additional detail such as who was present and where did the action take place. Once again, this is not an indicator of what truly occurred during the day, but appears to reflect the memory capacity of the respondent to remember more precise details. As for the rapport factor, the increased probability to change what was already reported, or that the interviewer appears to avoid using the pre-established codes to record the activities, seems to indicate that a richer conversational interaction was taking place, arguably because of a higher level of flexibility and rapport (understood as the construction of a friendly environment for the survey interaction) established from the beginning of the interaction between interviewers and respondents.

In any case, I believe that what may contribute to settling an emerging debate between estimating formative and reflective models centers on how researchers operationally define variables. In the survey field, for instance, several time-use researchers (e.g., Fricker and Tourangeau, 2010; Freedman et al., 2013) have argued that a larger number of reported activities constitute better data quality. However, measurement experts seem to strive for more precision in the language and the operationalization of constructs, and thus, what for survey researchers constitute data quality, for measurement researchers constitute the behaviors that enables the reporting of a larger number of activities. In this research I struggled with these two different perspectives, but I concluded that it is more likely that what is being measured are behaviors and not a characteristic of the data themselves. I arrived at this conclusion, both
because of the arguments presented by my committee members, but also because of the evidence that emerged from the estimated models. Therefore, for instance, I decided to modify the name of the constructs, and instead of just naming the factors as ‘measurement-related data quality factor’, and ‘nonresponse-related’ data quality factor, I called the first one motivation to respond factor, and the second one, memory-related factor.

**Prediction of motivation and memory from interview rapport.** After obtaining an acceptable three-factor CFA model for the two data quality factors (motivation and memory) and interview rapport, I estimated a full structural equation model predicting data quality from interview rapport. The estimation of the full structural model indicated that interview rapport significantly predicted both quality factors, though in opposite directions. Motivation to offer complete information, arguably the result of an optimizing behavior on the part of the respondent (Krosnick, 1991), is positively predicted by rapport. However, the memory capabilities to provide additional details of the reported activity was negatively predicted by interview rapport.

The finding that higher levels of rapport result in more motivated respondents is self-explanatory. In this case, rapport seems to encourage respondents not to satisfice (Krosnick, 1991) and/or to rush through the interview. In sum, one may say that rapport promotes a higher level of effort that enables the reporting of more complete answers. On the other hand, the apparently counterintuitive finding that rapport is detrimental to memory is supported by theory and previous research. For example, Belli et al. (2013) has shown that while rapport can be beneficial to the accuracy of retrospective reports in
some circumstances, it can also be detrimental in others, especially when sensitive topics are at stake; this seems to occur because a higher level of rapport may bias respondents into offering sociably desirable answers so as to not disappoint the interviewer. Rapport also has been related to interviewers obtaining higher levels of acquiescent reports (Olson and Bilgen, 2011), and to ‘speeding’ (Olson and Peytchev, 2007). In effect, because of rapport, interviewers and respondents make an effort to create a positive environment for their interaction by making efforts to defer to or not offend each other (Ross & Mirowsky, 1984). Nevertheless, precisely because of this rapport, the survey interaction may be negatively affected and more prone to measurement error. For example, because of rapport an interviewer may decide to change the wording of the question because it seems too long or intrusive, or skip the question altogether, and thus bias results. In the case of the 2010 ATUS and the results of the models I estimated, one may think that because of rapport, interviewers may be fearful of imposing on the respondent’s agreeableness towards the survey, and decide not to insist about asking about the additional details such as whether someone was present or where the activity took place, which could result in more who and where errors, especially during instances in which they observe that respondents are having difficulty remembering these details.

**Predictors of motivation, memory, and rapport.** Motivation, memory and rapport were predicted from certain variables, including demographics, social capital, and ‘busyness’. As for the influence of demographic variables, it was found that age positively predicted motivation, but negatively predicted data quality related to memory processes. It seems, therefore, that older people may be more motivated to provide a
more complete listing of their activities compared to younger respondents, but that compared to younger people, older people are also more cognitively challenged and are expected to provide less details about their reported activities. Gender was found to be a significant predictor of motivation, memory, and rapport. Compared to men, women are more motivated and have higher levels of rapport, indicating that women were more likely to make an effort to build rapport, as well as to report more activities from the previous day in comparison to men. Men, however, seem to have less problems remembering details about their activities compared to women. As for education, it was found that, overall, more educated respondents were more motivated and had more rapport; surprisingly, the more educated were less likely to provide additional details about their activities. One reason for this result is that more educated respondents may be more concerned about their privacy compared to less educated respondents.

Regarding the busyness variables, and consistent with the expectation, I found that compared to respondents out of the labor force, respondents in any type of occupation (i.e., busier) were less motivated and were less likely to build rapport. Finally, with regard to the social capital variables, these were significant positive predictors of motivation and rapport, but negative predictors of memory. One may assume that respondents with a higher experiential complexity (children in the household, more family members living in the same households), certain details about the activities are more easily forgotten than for respondents with a lower level of experiential complexity.

**Fidelity Analysis.** The results of the fidelity analysis were not as clear or conclusive as I would have expected. The inclusion of survey weights based in the
different factor scores did not produce meaningful differences across the different models estimated. To begin with, the variance explained for all models was not extremely impressive. However, it was found that models estimated with the memory factor weights did fare better in terms of variance explained in overall health. In effect, compared to models using motivation and rapport weights, the model estimated using the memory weights explained about 10% more of the variance in overall health. Also, the memory model was the only one in which one additional variable reached significance (number of children).

The reason why the different models did not display noticeable differences might be that the factor scores are fairly correlated among each other. It remains to be tested whether the combination of all three factors may produce different results. In any case, this ‘fidelity’ analysis is part of the exploratory effort of this dissertation, and I believe it has opened up many possibilities for further research, especially in terms of how to operationalize data quality and how to incorporate it in survey estimates.

**Limitations, Implications, and Future Directions for this Dissertation**

Among the main limitations of my research is that overall health is not necessarily the best proxy for wellbeing. Wellbeing is a much broader concept than that of overall health, which focuses on only one aspect of people’s wellbeing. In that sense, in order to extend the present research, it would be necessary to conduct similar analyses using data from the 2012 and 2013 ATUS wellbeing modules where overall wellbeing was actually measured, but for which I lacked the paradata variables. In any event, I believe that conducting a replicate study using the 2012 and 2013 ATUS data is feasible. Indeed, a
measurement model for ‘data quality’ can be estimated using exclusively public release data, and a fidelity analysis can still be conducted using the public release datasets and the resulting factor scores; the only analysis that will not be able to be replicated is the measurement model for *rapport* that is based on several paradata or process variables. In such a replicate analysis it will be interesting to see whether the reliability and rounding variables remain uninformative as was the case in this dissertation, and that was one of least expected results of my work.

Another issue with this dissertation may be the way factor scores were incorporated in the analyses. It is necessary to use a different method to explore the influence of data quality and interview rapport on analyses that use survey data. An option would be to incorporate factor scores not as weights, but as covariates that interact with other predictors. This is also a feasible study to be conducted in the future. Additionally, given the exploratory nature of this research, another option to investigate the influence of *motivation, memory,* and *rapport* factors would be the conduction of simulation studies, where the research is able to establish ‘truth’, thus allowing for a more clear understanding the influence of these factors in a more controlled setting.

An additional future direction of this work is a further consideration of the dimensionality of the three factors herein explored. In particular, it may be worthy to explore whether *motivation* and *rapport* belong to the same factor or not. For the moment, it has been established that they are in fact separate factors, one predicting the other. This separation is based on the fact that *motivation* is indicated by end-result variables (e.g., total number of activities reported in the public release dataset), whereas
rapport is indicated by process variables (audit trails). However, it is possible to further separate those process variables into, for instance, final missing who codes and final missing where codes from those missing who and missing where that would be eventually completed in the course of the interview. This lack of separation might be a source of confounding in the present research, which will need to be further addressed.

**Implications for Survey Research.** Some implications of the final conclusions of this dissertation are related to the influence of demographic variables on data quality identified in the analyses. First of all, age is an important predictor of data quality, though it functions differently depending on which data quality factor is being considered. It seems to be the case that older people are more willing to engage in fruitful and longer conversations in which potentially more activities will be reported. The longer interviews could, of course, be also the result of older people needing more time to process and report information, but at the end of the day, they seem to be more willing to give a more complete account of their day. Therefore, I believe it would be useful to train interviewers in this respect by devising strategies with which to encourage younger respondents to provide more thorough reports. On the other hand, given that older respondents tend to produce more memory errors, ATUS could devise some way to help older respondents to better remember their activities from the previous day, perhaps by combining the diary technique with some sort of memory aid that reminds participants of details that took place during the actual diary day.

Finally, as argued by Fricker and Tourangeau (2010), and as expected in my dissertation, it is important to note that social capital variables do seem to significantly
influence data quality related to the willingness to provide complete diary reports. Indeed, even though ATUS respondents who are married, who live in larger household sizes, and where young children are present may be taken as having less time to provide thorough responses, this type of respondents actually offer better and more complete diary reports. Thus, interviewers who are dealing with respondents with less social capital, may also help them to better respond by drawing on some type of strategy that would convey the importance of understanding how these respondents use their time.

A final implication is that the evaluation of data quality and error in surveys benefits greatly by the more widespread use of psychometric techniques such as that of structural equation modeling. Indeed, with all the statistical techniques available, survey researchers should not continue analyzing data as if survey responses were completely reliable, and should utilize techniques such as SEM that can handle measurement error in a more sophisticated manner.
REFERENCES


APPENDIX A
MPLUS SYNTAX FOR FINAL CFA MODEL FOR DATA QUALITY

TITLE: DATA QUALITY MODEL 6
!Model2redo: Rounding rate and reliability removed;
!Model3redo: 2 factors;
!Model4redo: Where error WITH Who error;
!Model5redo: miss_r WITH activ_tr;

DATA:
FILE IS ATUS_SEM_unf.csv;

VARIABLE:
NAMES ARE Audittr TUCASEID tot_entr Tot_time SumPromp
TotActPR Sum_sec E_whr_ch E_Act_ch Emisswho Emisswhr
Dmisswho Dmisswhr E_insert E_delete D_insert D_delete
Everbat Dverbat hours_p TUAVGUR TUCPSDP TUINCENT INTDQUAL
English tot_call tot_week D_REFUS no_sleep no_groom no_eat
misbasic tot_miss E_round round_rt e_any_er D_error
GENDER_M under6 Cooperate HETENURE FAM_INC S_Emp_Ad hispanic
educa lessHS highsch somecoll bachelor graddeg maritdeg
CHILDNUM NUMHOU SPPRES Employed preschil occup
e_reliab reliabrat m_happy m_meanin m_pain m_sad m_stress
m_tired ovhealth HBP_HYP PAINMED rested age47 INT_ID
entries_r duration totact_r second_r act_ch_r whr_ch_r
round_r eround_r reliab_r miss_r error_r misswhor misswhrr
e_miss_r nev_ref promp_tr prompts entr_tr entr_cat entries
act_ch_tr actchcat activ_ch whr_ch_tr whrchcat where_ch
sec_tr second verb_tr verb_cat verbatim minut_tr time_cat
minutes activ_tr act_cat totalact Execut Service Support
NILF vrb_tr_r prom_trr sec_rev;

USEVARIABLES ARE Audittr INT_ID activ_tr miss_r misswhor
misswhrr minut_tr error_r;

IDVARIABLE IS AuditTr;
CLUSTER IS INT_ID;
CATEGORICAL ARE miss_r misswhor misswhrr error_r;
MISSING ARE ALL (-9999);

ANALYSIS:
TYPE=COMPLEX;
ESTIMATOR IS WLSMV;
PARAMETERIZATION IS THETA;
!DIFFTEST=Quality5.dat;

SAVEDATA:
DIFFTEST=Quality6.dat; SAVE = FScores;
FILE = Quality6THETAS.dat;

PLOT:
TYPE = PLOT1 PLOT2 PLOT3;

MODEL:
QUALITY1 BY activ_tr* miss_r* minut_tr*;
QUALITY2 BY misswhor* misswhrr* error_r*;
QUALITY1@1;
QUALITY2@1;
misswhor WITH misswhrr;
miss_r WITH activ_tr;
APPENDIX B
MPLUS SYNTAX FOR FINAL CFA MODEL FOR INTERVIEW RAPPORT

```
TITLE: Interviewer Rapport - MODEL 2;
DATA: FILE IS ATUS_SEM_unf.csv;
VARIABLE: NAMES ARE Audittr TUCASEID tot_entr Tot_time SumPromp
       TotActPR Sum_sec E_whr_ch E_Act_ch Emisswho Emisswhr
       Dmisswho Dmisswhr E_insert E_delete D_insert D_delete
       Everbat Dverbat hours_p TUAVGDUR TUCPSDP TUINCENT INTDQUAL
       English tot_call tot_week D_REFUS no_sleep no_groom no_eat
       misbasic tot_miss E_Round round_rt e_any_er D_error
       GENDER_M under6 Cooprare HETENURE FAM_INC S_Emp_Ad hispanic
       educa lessHS highsch somecoll bachelor graddeg marital
       CHILDNUM NUMHOU SPPRES Employed preschil occup
       e_reliab reliabrt m_happy m_meanin m_pain m_sad m_stress
       m_tired ovhealth HBP_HYP PAÎNMED rested age47 INT_ID
       entries_r duration totact_r second_r act_ch_r whr_ch_r
       round_r eround_r reliab_r miss_r error_r misswhor misswhrr
       e_miss_r nev_ref promp_tr prompts entr_tr entr_cat entries
       act_ch_tr actchcat activ_ch whr_ch_tr whrchcat where_ch
       sec_tr second verb_tr verb_cat verbatim minut_tr time_cat
       minutes activ_tr act_cat totalact Execut Service Support
       NILF vrb_tr_r prom_trr sec_rev;

USEVARIABLES ARE Audittr INT_ID
       entries sec_tr activ_ch where_ch verb_tr;

IDVARIABLE IS AuditTr;
CLUSTER IS INT_ID;
MISSING ARE ALL (-9999);

ANALYSIS:
       TYPE=COMPLEX;
       ESTIMATOR IS MLR;

SAVEDATA:
       SAVE = FScores;
       FILE = Intervredo2MLR.dat;

PLOT:
       TYPE = PLOT1 PLOT2 PLOT3;

OUTPUT:
       MODINDICES (3.84)
       STBYX
       RESIDUAL

MODEL:
       INTERV BY entries* sec_tr* activ_ch* where_ch* verb_tr*;
       INTERV@1;
       ACTIV_CH WITH SEC_TR;
```
APPENDIX C
MPLUS SYNTAX FOR FINAL THREE-FACTOR MODEL (TWO-STEP APPROACH)

TITLE: Two-step Model 5
!Model 2: Drop entries
!Model 3: ACTIV CH WITH SEC_tr;
!Model 4: miss_r WITH activ_tr;

DATA: FILE IS ATUS_SEM_unf.csv;

VARIABLE: NAMES ARE AuditTr TUCASEID tot_entr Tot_time SumPromp TotActPR Sum_sec E_whr_ch E_Act_ch Emisswho Emisswhr Dmisswho Dmisswhr E_insert E_delete D_insert D_delete Everbat Dverbat hours_p TUAVGDUR TUCPSDP TUINCENT INTDQUAL English tot_call tot_week D_REFUS no_sleep no_groom no_eat misbasic tot_miss E_round round_rt e_any_er D_error GENDER_M under6 Cooperate HETENURE FAM_INC S_Emp_Ad hispanic educa lessHS highsch somecoll bachelor graddeg marital CHILDS NUMHOU SPPRES Employed presch occupe_reliab reliabrt m_happy m_meani m_sad m_stress m_tired ovhealth HBPHYP PAINMED rested age47 INT_ID entries_r duration totact_r second_r act_ch_r whr_ch_r round_r eround_r reliabr_r miss_r error_r misswhor misswhrr e_miss_r nev_ref promp_tr prompts entr_tr entr_cat entries act_ch_tr actchcat activ_ch whr_ch_tr whrchcat where_ch sec_tr second verb_tr verb_cat Verbatim minut_tr time_cat minutes activ_tr act_cat totalact Execut Service Support NILF vrb_tr_r prom_trr sec_rev;

USEVARIABLES ARE AuditTr INT_ID activ_tr miss_r misswhor misswhrr minut_tr error_r sec_tr activ_ch where_ch verb_tr;

IDVARIABLE IS AuditTr;
CLUSTER IS INT_ID;
CATEGORICAL ARE miss_r misswhor misswhrr error_r;
MISSING ARE ALL (-9999);

ANALYSIS: TYPE=COMPLEX;
ESTIMATOR IS WLSMV; PARAMETERIZATION IS THETA;

SAVEDATA: DIFFTTEST=Twostep3.dat;
SAVE = FScores;
FILE = TwoStep3THETAS.dat;

PLOT: TYPE = PLOT1 PLOT2 PLOT3;

OUTPUT: MODINDICES (3.84)
STDYX RESIDUAL

MODEL: QUALITY1 BY activ_tr* miss_r* minut_tr*;
QUALITY2 BY misswhor* misswhrr* error_r*;
INTERV BY sec_tr* activ_ch* where_ch* verb_tr*;
QUALITY1@1;
QUALITY2@1;
INTERV@1;
miss_r WITH activ_tr;
ACTIV_CH WITH SEC_tr;
APPENDIX D
MPLUS SYNTAX FOR FINAL SEM MODEL

TITLE: SEM Model 1
DATA: FILE IS ATUS_SEM_unf.csv;
VARIABLE: NAMES ARE Audittr TUCASEID tot_entnr Tot_time SumPromp TotActPR Sum_sec E_whr_ch E_Act_ch Emisswho Emisswhr Dmisswho Dmisswhr E_insert E_delete D_insert D_delete Everbat Dverbat hours_p TUAVGDUR TUCPSDP TUINCENT INTDQUAL English tot_call tot_week D_REFUS no_sleep no_groom no_eat misbasic tot_miss E_round round_rt e_any_er D_error GENDER_M under6 Cooperate HETENURE FAM_INC S_Eemp_Ad hispanic educa lessHS highsch somecoll bachelor graddeg marital CHILDCN NUMHOU SPPRES Employed preschool occup e_relib ralliabrt m_happy m_meanin m_pain m_sad m_stress m_tired ovhealth HBP_HYP PAINMED rested age47 INT_ID entries_r duration totact_r second_r act_ch_r whr_ch_r round_r eround_r reliab_r miss_r error_r misswho misswhrr e_miss_r nev_ref promp_tr prompts entr_tr entr_cat entries act_ch_tr actchcat activ_ch whr_ch_tr whrchcat where_ch sec_tr second verb_tr verb_cat verbatim minut_tr time_cat minutes activ_tr act_cat totalact Execut Service Support NILF vrbl_tr_r prom_trr sec_rev;

USEVARIABLES ARE Audittr INT_ID activ_tr miss_r misswhor misswhrr minut_tr error_r sec_tr activ_ch where_ch verb_tr;

IDVARIABLE IS AuditTr;
CLUSTER IS INT_ID;
CATEGORICAL ARE miss_r misswhor misswhrr error_r;
MISSING ARE ALL (-9999);

ANALYSIS: TYPE=COMPLEX;
ESTIMATOR IS WLSMV; PARAMETERIZATION IS THETA;
SAVEDATA: DIFFTEST=SEM1REDO.dat; SAVE = FSOCORES;
FILE = SEM1REDOTHETAS.dat;
PLOT: TYPE = PLOT1 PLOT2 PLOT3;
OUTPUT: MODINDICES (3.84)
STDYX RESIDUAL

MODEL: QUALITY1 BY activ_tr@1 miss_r* minut_tr*;
QUALITY2 BY misswhor@1 misswhrr* error_r*;
INTERV BY verb_tr@1 sec_tr* activ_ch* where_ch*;
ACTIV_CH WITH SEC_TR;
miss_r WITH activ_tr;
QUALITY1 ON INTERV;
QUALITY2 ON INTERV;