ON THE RELEVANCE OF THE REFERENCE PERIOD IN YOUTH MENTAL HEALTH OUTCOME QUESTIONNAIRES

By

Manuel Riemer

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

December, 2006

Nashville, Tennessee

Approved:

Professor Leonard Bickman
Professor David S. Cordray
Professor Thomas M. Smith
Professor David Cole
Professor Georgine Pion

Copyright © 2006 by Manuel Riemer
All Rights Reserved

DEDICATION

I dedicate this dissertation to my parents Monika and Willi Riemer, who gave me the support and the space to grow as a person as well as a scholar.

ACKNOWLEDGEMENTS

I would like to thank several people who supported me throughout this journey. First, I would like to thank my advisor, Len Bickman, for always treating me as a colleague and for providing me with many opportunities to develop my skills as a scholar and researcher. The kind of working relationship we were able to establish has resulted in some very interesting and productive years. While it is strange to think that I will not be his student anymore, I am sure that we will continue collaborating on many projects in the future. Second, I would also like to thank the members of my committee for their support and helpful feedback along the way.

Next, I would like to thank my friend Lynne Wighton who played an important role in my development as a scholar. With her amazing skills as an editor she taught me, as a foreigner, how to become proficient in writing academic papers in the English language. I am also very thankful for her belief in my academic abilities, which I have always found to be very encouraging.

I would also like to thank Isaac Prilleltensky for his intellectual mentorship and his friendship. I thank Warren Lambert for his statistical advice over the years. Furthermore, I would like to thank my friends whose love and support means a lot to me. I thank Martina Preisler, Stephan Koch, Ron Kastner, Stephanie Reich, Vicky Ngo, Doug Morse, Jeff Nyquist, and Kelly Richardson among many others.

Finally, I would like to thank my parents and my girlfriend Jocelyn. My parents gave me unconditional love and support throughout my life, for which I am very grateful. They provided me with the base that allowed me to explore the world. Jocelyn's love and support during these last years has been very important to provide me with a positive balance, especially when things were difficult and seemed insurmountable.

Dankeschön!

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

LITERATURE REVIEW

Introduction

In recent decades, the mental health field began to realize what other service providers and the industrial sector had already discovered a while ago: in order to assure quality and effectiveness of services, one needs to employ methods of quality assurance and continuous quality improvement. This movement, however, poses some new challenges to the mental health field. Quality assurance and quality improvement require ongoing monitoring of outcomes. However, the definition and measurement of outcomes in mental health is not straightforward. Comprehensive diagnostic interviews administered by clinical experts that have been used in the past to assess mental health status are not a feasible option for frequent and ongoing assessment in practical settings and for administration in large-scale outcome studies (Kessler, Wittchen, Abelson, & Zhao, 2000). Consequently, the field has to rely mainly on self-report or, in the case of children and youth, also on the report by caregivers. Self-reports, however, can be unreliable because of measurement error and the possibility of unconscious and conscious bias in the person reporting (Baldwin, 2000). If these biases and errors become large enough, they threaten the validity of the inferences that are drawn from the outcome instruments. Much of the developmental work in regard to youth mental health outcome

scales has been focused on the substantive content with too little attention to the methodological problems associated with self-report. This dissertation will address this problem by focusing on one specific methodological aspect of self-report in the context of youth outcome measures—the time or reference period the respondent is asked to consider when answering questions.

I will investigate this question using the Symptom and Functioning Severity Scale (SFSS) that is being used regularly to assess and monitor clinical outcomes in practical settings. This question of what reference period is appropriate arose in the context of our research at the Center for Evaluation and Program Improvement at Vanderbilt University. The answer to this question, however, has implications beyond our own research as I will discuss later in this paper. At the Center, we were faced with the challenge of developing a battery of measures to be used in the context of a comprehensive feedback system for clinicians in community mental health settings (Bickman, Breda, Dew, Lambert, Pinkard, Riemer, et al., 2006), one of which was the SFSS. The SFSS covers a range of common internalizing and externalizing problems in youth with items based on four of the most prevalent childhood disorders: attention deficit hyperactivity disorder (ADHD), conduct/oppositional disorder, depression, and anxiety. In addition, it includes items related to peer and family relationship problems. The goal was to capture the general mental health status of the child as it develops over the course of treatment with only 33 items. Respondents are asked to rate each on item on a five-point Likert type scale (*Never, Hardly Ever, Sometimes, Often, Very Often*).

One of the questions that we had during the development process was how often this instrument should be administered (weekly, every other week, once a month, or

maybe every six months)? Related to this question was what the reference period stated in the instructions of the questionnaire should be. Should it be two weeks simply because that is the frequency of administration that we decided on? Also, should there be a different version for intake (e.g., asking about the last six months) than during the treatment phase (e.g., two weeks)? If we use a different reference period for intake, what reference period should be used at discharge: the same as the one at intake to have comparable pre-post assessments or the same as the treatment phase to prevent overlap in time? These questions depend on the answer to the question of whether using different reference periods would change clients' and their caregivers' response behaviors.

This paper is an attempt to find an answer to this question about the relevance of the reference period in regard to the response behavior of youths and caregivers completing a youth mental health outcome measure. In this introductory chapter I will begin with a discussion of mental health outcome measurement in general and measurement in youth mental health in particular to provide the reader with an idea of some of the challenges this field has faced and how the question of the reference period has been handled so far. I will then review the relevant literature on the survey response process in general and the reference period in particular. I will illustrate that according to this literature, the reference period in measures like the SFSS should, on average, not make a meaningful difference in the way people respond. I will end the introduction with four hypotheses that will allow me to test this assumption of no difference in several aspects. In the second chapter I will describe the procedures and methods used for the current study. This includes a psychometric analysis of the SFSS using a Rasch modeling approach. The third chapter provides an in-depth discussion of the analytical approach

that allows me to test hypotheses of no difference. This is followed by the specifics of the analysis for each hypothesis test and the presentation of the results. I will conclude with a discussion of the results, the consequences of my findings for the field of mental health, the limitations of the presented research, and some suggestions for future research.

Outcome Measurement in Youth Mental Health

The mental health sector offers a variety of services for youth such as inpatient and partial hospitalization, residential care, therapeutic foster care, day treatment, outpatient services as well as in-home and outreach services. As part of these services, an individual or group of providers undertake referral, intake, diagnostic evaluation and formulation, collaborative treatment planning, implementation of treatment, and termination (Bickman, Nurcombe, Townsend, Belle, Schut, & Karver, 1998). The type of outcome measurement that is discussed in this dissertation concerns primarily the assessment of success in regard to the implementation of treatment. That is, it assesses whether the treatment intervention led to an improvement in the youth's mental health status. While some diagnostic instruments are also being used to monitor outcomes, diagnosis is generally distinguished from outcome evaluation and quality improvement and requires different qualities from a measurement instrument (Sperry, Brill, Howard, Grissom, 1996).

While it is clear that the primary purpose of outcome measurement and monitoring is to improve the efficiency and effectiveness of mental health services, not all stakeholders agree on what aspects of outcomes should be measured. Outcomes can be assessed at different levels: consumer, clinician, treatment, clinic, and overall system. Content areas that have been discussed for potential monitoring include the severity and

acuity of symptoms, the functional impairment and strength of the youth, family functioning, quality of life, consumer satisfaction, the goals of treatment, readiness for change, the quality of the therapeutic alliance, and adherence to treatment (Bickman, et al., 1998). The SFSS was developed to assess symptom severity as well as functional impairment at the consumer level. The reasons for focusing our investigation on these types of outcomes do not reflect a value judgment about which outcomes are the most important ones, but rather reflected practical as well as methodological concerns. In fact, the SFSS instrument is part of a comprehensive measurement package, the Peabody Treatment Measurement Battery, developed by Leonard Bickman and his team (Bickman, et al., 2006). This battery includes scales for the assessment of treatment motivation, life satisfaction, caregiver strain, and therapeutic alliance, among others. Measures for symptoms and functioning are more common, however. Thus, it is reasonable to focus the investigation of the relevance of the reference period for this type of measure.

Multiple instruments for assessing symptom severity and functioning of youths exists. Among the most prominent and prevalent are: the three measures developed by Achenbach, which include the Child Behavior Checklist (CBCL, Achenbach, 1991a), the Youth Self Report (YSR, Achenbach, 1991b), and the Teacher Rating Form (TRF; Achenbach, 1991c); the Child and Adolescent Functional Assessment Scale (CAFAS; Hodges, 1990); the Youth Outcome Questionnaire and Self-Report Youth Outcome Questionnaire (YOQ and SR-YOQ; Burlingame, Wells, Cox, Lambert, Latkowski, Ferre, 2005; Wells, Burlingame, & Rose, 1999); the Strength and Difficulty Questionnaire (SDQ; Goodman, 2001); and the Ohio Youth Problem, Functioning, and Satisfaction

Scales – Short Form (Ohio Scales, Ohio Department of Mental Health, 2004). These measures all overlap in content and purpose even though they may emphasize different aspects of symptom severity and functioning and may cover additional content areas (e.g., hopefulness). However, there are some substantial differences that are pertinent to the present study: (a) type of informant; (b) number of items; (c) the metric the score is reported on; and (d) the time period (i.e., the reference period) the respondents are asked to recall. In order to compare inferences about treatment efficiency and effectiveness that result from these outcome measures, it is important to understand whether these differences are meaningful and how they manifest themselves.

Type of Respondent

The types of informants used for these measures include the youth, caregivers, clinicians, and teachers. There is clear evidence that the type of informant matters. In a meta-analysis by Achenbach and his colleagues (Achenbach, McConaughy, & Howell, 1987), for example, they found that while the agreement between similar informants (e.g., two parents) is relatively high (mean r=.60), the correlation between different types of informants (e.g., parent and teacher) is low (.28) and even lower if the scores of the child or youth are compared to the those of others (.22). A more recent review by Meyer and colleagues (Meyer, Finn, Eyde, et al., 2001) generally confirms these findings. Thus, it is clear that the respondent type needs to be considered when interpreting and comparing the scores of different outcome measures. If, for example, one treatment's effectiveness was assessed using the CAFAS (clinician-based ratings) and another treatment's effectiveness was assessed using the YSR, differences in the findings may be simply due to the use of different respondents. In this dissertation I will investigate two

types of respondents: the youth and the caregiver. However, I will conduct the analysis separately for each respondent type.

Number of Items

A systematic review that investigates the comparability of instruments with significantly different numbers of items does not exist to my knowledge. However, there is some evidence from validity studies that compared longer instruments to shorter ones using the same informants. For example, the correlation of the SDQ (caregiver version) total score (25 items) with the CBCL total problem score (118 items) was .86 (Goodman & Scott; 1999). The developers of the Ohio Scales validated the current short form with a previous longer version. The Ohio Scales Problem Severity Scale was correlated at .80 with its longer predecessor, and for the Functioning Scale this correlation was .91 (Ohio Department of Mental Health, 2004). These high correlations, often in the same range as the internal reliability estimates of these measures, seem to suggest that the number of items may not be an issue when comparing inferences from different instruments as long as the items are comparable overall. However, a more systematic investigation into this question may be warranted. For the purpose of the current investigation I will use the same instruments for all comparisons keeping the length of the measure constant.

Metric of Score

In a recent paper Blanton and Jaccard (2006) drew our attention to the problem of arbitrary metrics in psychology. According to these authors, the term metric *"*refers to the numbers that the observed measures take on when describing individuals' standings on the construct of interest." (p. 27) Many of these metrics are arbitrary, that is, the function

describing the relationship of individuals' true score on the latent construct of interest to their observed score on the response metric and the parameter values of that function are unknown. As an illustration consider height expressed in centimeters as compared to a score on the YOQ. If a person's height is 200 cm, we have a pretty good idea that we are dealing with a pretty tall person and that a person of 100 cm is half that size (of course, this assumes familiarity with the metric system). A score of 80 on the YOQ, however, does not provide one with a comparable idea of what it represents in regard to the youth's mental health status and in no way can one infer that a person with a score of 40 has half the severity level.

One case example Blanton and Jaccard discuss are measures used to evaluate the real-world importance of clinical interventions similar to those discussed in this dissertation. A common strategy for these types of measures is to obtain a norm sample and then present current scores and changes in scores in standard deviation units of that norm group. However, "examining scores in terms of standard deviation units is simply a rescaling of the metric and does not make the metric any less arbitrary. There is no sense of how much the underlying psychological construct has changed when someone's standard score of 2.2 is reduced to a standard score of 1.8, nor is it known if there are any implications of that change for the individual being treated." (Blanton and Jaccard, 2006, p.37) These authors see a more promising approach to be a cut-off score defined as whether the client is statistically more likely to be considered dysfunctional than functional. However, this approach is not much discussed by these authors. The issues with arbitrary metrics are specifically relevant if one wants to compare the scores from different scales supposedly measuring the same construct. Each of the existing youth

mental health outcome scales discussed earlier uses a different scaling. Most often, their scores are based on the average or sum of the raw item ratings. In this dissertation, I will take this issue of arbitrary metrics into account in two regards. First, I will transfer the raw scores of the SFSS into a measure score using a Rasch model approach. One of the biggest advantages of this approach is that Rasch measure score is on a true interval scale level. This will allow me to evaluate the properties of the SFSS as a true scale and use interval level scores in most of the analyses. Secondly, I will take advantage of the manual of SFSS (Bickman et al., 2006) providing a cut-off score for cases considered being in the high-severity range, which is comparable to the clinical range in the CBCL and YSR (e.g., Achenbach, 1991a, 1991b).

Reference Period

Finally, another difference among these youth mental health outcome measures is the reference periods that these instruments ask the respondents to cover when recalling behaviors, emotions, and cognitions relevant to symptomology and functioning. The reference periods range from one week to six months (see Table 1). While all of these instruments state a reference period, some of them handle it rather flexibly. The instructions of the CBCL, YSR, and TRF, for example, ask the respondents to report about their experience "now or within the last 6 months" leaving it to the respondents to determine how far back they search their memory. The CAFAS gives the trained rater a choice of one month, 3 months, or whatever time frame seems appropriate to the rater completing the form. The SDQ asks about the last six months, but also provides a follow-up version that is the same as the six-month version except it uses a one-month reference period.

Table 1: Characteristics of Youth Mental Health Outcome Measures

Instrument	Respondent	Number of Items	Scoring*	Reference Period(s)
CBCL	Caregiver	118	Rating: 3-point Likert Score Range: 23-100	"now or within the last six months"
YSR	Youth	112	Rating: 3-point Likert Score Range: 23-100	
TRF	Teacher	118	Rating: 3-point Likert Score Range: 23-100	
CAFAS	Clinician	200	Rating: 0-30 Score Range: 0 - 240	"Last Month Last 3 Months Other ____"
SR-YOQ / YOQ	Youth Caregiver	64	Rating: 5-point Likert Score Range: -16 - 240	"during the past 7 DAYS"
SDQ	Youth Caregiver Clinician	25	Rating: 3-point Likert Score Range: 0 - 40	"over the last six month" OR "over the last month"
Ohio Scales	Youth Caregiver Clinician	48 48 40	Rating: 6-point & 5-point Likert Score Range: 0-120 (Symptoms) Score Range: 0-100 (Functioning)	"in the past 30 days"
SFSS	Youth Caregiver Clinician	33	Rating: 5-point Likert Score Range: 32 - 92 (Youth) 42 – 94 (Car. & Cl.)	"Over the last two weeks"

** Only the total scores are reported here*

According to Schaeffer and Presser (2003), "the choice of reference period is usually determined by the periodicity of the target events, how memorable or patterned the events are likely to be, and the analytic goals of the survey." (p. 71) Given the wide range in reference periods, this would imply that the developers of the discussed outcome measures have different ideas about these issues. Achenbach (1985) is probably the most explicit about his choice for a rather long reference period: "For easily observed problem behaviors of high frequency, assessment may need to span only a few weeks to provide a stable baseline. For problems that require more inference, are of low frequency, or comprise syndromes of covarying features, however, longer spans are needed to provide reliable and valid baselines." (p.145) However, users of the CBCL often administer the CBCL in shorter frequencies than six months. In a study by Henggeler and colleagues

(Henggeler, Rowland, Halliday-Boykins, et al., 2003), for example, they used the CBCL in time intervals of three months. Service providers I have been working with also administer the CBCL every three months.

The developers of the YOQ explain that without a clinical intervention the YOQ scores remain relatively constant over a short period of time (two to four weeks) but are sensitive to change in that time period if the youth is treated effectively (Burlingame, et al., 2005). They recommend a frequent administration (weekly or bi-weekly) because frequent repeated measurement would increase the reliability of growth rate data and, thus, provide a more sensitive tracking process.

The Ohio Scales, which also use a one-week reference period, have been developed as part of a state-wide outcome monitoring system. In this context the scales are administered every six months in the first year of treatment and annually thereafter. However, the authors of the Ohio Scale state that "the easy administration of the Ohio Scales allows the instrument to be used as frequently as the clinician would like." (Ohio Department of Mental Health, 2004; p. 8-15) However, neither the manual for the Ohio Scales nor the one for the YOQ discuss why they specified a reference period of one week. As can be seen in the use of the reference period of the CAFAS (one month, three months, other; see Table 1) it is also left to the clinician or the leadership of the organization using the CAFAS to determine what the actual reference period and frequency of administration should be.

These examples illustrate that there is no agreement in the field on how important an exact reference period is, whether it really matters, and if it does, what it should be. To my knowledge no systematic investigation of the impact of different reference periods on

mental health outcome measures for youth has been conducted. Thus, the only way for us to find answers to the question on the relevance of the reference period was investigating it ourselves. We were fortunate enough to work with a large mental health service provider that was willing to integrate a randomized experiment for this purpose into an already planned psychometric evaluation of the Peabody Treatment Progress Battery they were planning to use for their quality improvement project, Contextualized Feedback Intervention and Training (CFIT). In this experiment respondents were randomly selected to either complete a version of the SFSS with a two week reference period (referred to as the reference group) or a longer reference period (e.g., three months or six months; referred to as the comparison group).

It is important to note that the intent of the current study is primarily methodological in that I am testing whether the reference period in youth mental health outcome measures is an important stimulus to the respondent that would have an effect large enough to be detected in mental health outcome studies. This study was not designed to answer the substantive question of how people describe their problems over the period of two weeks compared to three or six months. However, this methodological study has important implications for this more substantive question. If youths and caregivers use the reference period as an important clue in generating their response choice, then the substantive question would be important to consider in the interpretation of these types of outcome measures. However, if the reference period is not a meaningful part of the questionnaire that affects the final choice of the answer, then the substantive question is of less relevance because respondents will use whatever reference period makes intuitive sense to them without paying much attention to the reference period

stated in the directions. The person interpreting the results of the questionnaire would have no way of telling what time period the youth or caregiver is referring to when they selected their answers to each question. In fact, it would be questionable to assume that just because one questionnaire used a six month reference period, the youth was reporting for the last six months while to infer that the next time they report about just the last two weeks because the reference period instruction was changed to two weeks. It seems that the relevance of the reference period has been underestimated in the context of developing youth outcome measures.

Because of the pressure to demonstrate quantifiable outcomes in providing mental health services, the mental health field is relying heavily on self-report measures. However, the response process in the context of self-report measures or proxy reports is quite complex and it is important that we pay more attention to the methodological implication of relying primarily on these types of measures. This current study is exploring just one aspect of this process among many others that we still need to understand better in order to make appropriate inferences about the mental health status of youths. I hope, however, that with this study I am raising the awareness that we need to pay attention to these types of details when developing outcome measures.

A review of the existing general literature on the cognitive aspects of the survey response process as well as previous studies on the relevance of the reference periods in other contexts will provide some insight into the complexity of the response process and help me determine whether I should or should not expect meaningful differences between the two groups.

The Reference Period in the Context of the Survey Response Process

Kessler and his colleagues (Kessler, et al., 2000) as well as Shiffman (2000) stress the importance of considering the cognitive aspects of survey responses especially in regard to autobiographical memory in the context of self-reports about behaviors related to psychiatric disorders. The research on the cognitive aspects of the survey response process has made great progress in the last twenty years and has significantly facilitated the understanding of how people answer standardized questions. Nevertheless, many questions remain.

Jobe and Herrman (1996) and Tourangeau, Rips, and Rasinski (2000) provide good overviews of the most prominent models of survey cognition. All of these models describe at least four stages in the response process: (a) comprehension of the instructions and the item, (b) retrieval of relevant information from the memory, (c) judgment, and (d) selection and report of the final response. Some of these models differentiate one or more of these stages further. The Information Exchange Theory by Mullin and colleagues, for example, emphasizes the process of question interpretation over and beyond the question comprehension process (e.g., Sander, Conrad, Mullin, & Herrmann, 1992). Also, it is not expected in these models that the respondent always passes through all of these stages or always does so sequentially. Dividing the process into several theoretical stages simply facilitates the discussion of the different relevant aspects of the survey response process. For the discussion of this process with regard to the potential impact of different reference periods, I will utilize the popular model by Tourangeau, which describes the process using the four stages described above (Tourangeau, 1984; see also Tourangeau et al., 2000). I will use this model in conjunction with the existing literature on the impact of different reference periods, which has mainly focused on the accuracy of recalling

events. I will apply this knowledge directly to the SFSS and the current study to determine whether one would expect differences between the reference group and the comparison group.

Comprehension

A typical outcome measure consists of some sort of instructions, a number of questions or statements, and two or more answer prompts. The comprehension of these elements by respondents is prone to error, often because the standard rules of conversation do not apply (Kessler, et al., 2000). In many cases the respondents do not understand a question the way it was intended by the instrument developers. In one of the first studies that investigated this issue systematically, Belson (1981) found that more than 70% of respondents interpreted at least some questions differently from the researcher. Subsequently much effort in the survey research field has been devoted to developing better questions and instructions (for comprehensive reviews see Sudman & Bredburn, 1982; Sudman, Bradburn, & Schwarz, 1996; Torangeau, et al., 2000). However, many difficulties remain and not all outcome measure developers appear to incorporate this accumulated knowledge about how to ask questions into full account.

The differences in the interpretation of a standardized question are best understood if we consider the mental representation of the question. Based on Rips (1995), Tourangeau and colleagues (Torangeau, et al., 2000) differentiate between the *representation of the sentence* and the *representation about the sentence*. The former "consists of a specification of the underlying grammatical and logical structure of the sentence, together with the lexical representation of the individual words it contains." (p.31). The representation about the sentence "consists largely of inferences that the

interpreter draws from the sentence in conjunction with other knowledge that he or she has available on that occasion." (p.31) Thus, the latter is dependent on characteristics of the respondents (e.g., level of knowledge, their perspective), knowledge of the question author, the situation, and the context. For example, a youth may have no problem understanding the words and the grammar of the question "Did you argue with adults in the last 2 weeks?" but could have a very different view of what to include in the category of an argument than what the person developing or evaluating the question has in mind. While it can be argued whether the authors of the different outcome measures for youth always accomplished their goal to write the questions in a way that is appropriate for the target audience, in the context of this paper I will not pay much attention to the *representation of* the question. Especially in regard to the reference period, it is safe to assume that youth will be able to understand the words "two weeks" just as well as "six months." More ambiguity, however, is likely to exist in regard to the *representation about* the items.

Ambiguity and vagueness of the items can lead respondents to interpret the questions in variable ways (Tourangeau, et al., 2000). A consequence of this is that the interpretation of the question can be influenced by factors other than the actual content such as the formal characteristics of the question (Schwarz, Strack, Müller, & Chassein, 1988). Schwarz et al., found that respondents interpreted the question of how often they felt "really annoyed" differently, if different frequency scales were used. In the low frequency scale condition (less than once a year to more than once every 3 months) people reported more extreme types of situations compared to those in the high frequency condition (less than twice a day to several times a day). Similarly, the reference period

has been found to influence respondents' interpretation of the meaning of certain questions.

There are several studies I found that are relevant here. In a couple of studies by Thomas and Diener (1995) on the recall accuracy of positive and negative emotions (both in regard to frequency and intensity) concurrent versus retrospective reports were examined. In the concurrent condition in their first study, 40 undergraduate students were asked to report the frequency and intensity of experienced emotions at four random times a day; the retrospective report was obtained after three weeks. In the second study with 103 students the respective time frames were once at the end of the day (concurrent) and six weeks (retrospective). Thus, indirectly they introduced different reference periods for the recall to their respondents. What they found was that the ratings of negative as well as positive intensity were significantly higher in the retrospective report than in the current reports which had a much shorter reference period. These authors suggest that "Possibly, subjects used a different scaling metric for momentary and daily emotions than longer time periods, rather than simply recalling more intense time." (p. 295) However, these studies were not designed to test this question directly.

In other studies, the impact of the reference period on how subjects interpret a question was investigated directly. In the first study by Winkielmann, Knäuper, and Schwarz (1998), 111 undergraduate students interpreted the question of how often they got angry differently when asked about the last week compared to the last year. The results indicate that in the former case the respondents reported less intense and more frequent episodes of anger while for the one year reference period they reported more extreme cases. In another experiment by Igou, Bless, and Schwarz (2000) with 177

German students, this finding was reproduced with a one day and a six month reference period. In a second experiment with 97 undergraduate students reported by Winkielmann et al. (1998), they manipulated the level of ambiguity in the question. In one condition they left the term "anger" undefined, while in the other condition they provided a definition for that term. In the former condition, the results replicated the findings of the two studies mentioned above. In the second condition, however, the difference in the type of anger situations the students reported disappeared. The consequence was that in the former case, the students did not report a higher frequency of anger situations using vague quantifiers (i.e., a 9-point response scale from *hardly ever* to *very frequently*) for the one year reference period compared to the one week reference period while in the latter case they did. In a third experiment with 92 students, Winkielmann et al. (1998) found that respondents reported different types of anger events when the six month question followed the same question with a one week reference period compared to the reversed order.

The findings referenced so far would suggest that for the short (two week) and the long (3 months / 6 months) reference period in the SFSS one should expect a difference in the frequency reports for items that are very well defined and less of a difference in items that are ambiguous and vague. However, these studies have something in common that is important to consider. In the studies referenced above, respondents were only asked one or two questions. As a consequence, the reference period was salient to the respondent because it was unique to the question. What would happen if the reference period is not unique to the question, but instead is used for multiple questions? Igou et al., (2002) tested this in two experiments. In the first experiment they provided the students

with either a question about the frequency of anger alone or they presented four different types of questions with the anger question being in either third or fourth position. The two different reference periods ("today" and "the past six months") were part of an introduction that preceded the questions. What they found was that when the reference period pertained only to the anger question, the students provided less extreme examples for the short reference period compared to the longer one. However, when the same reference period referred to multiple questions the difference was not significant. To test whether this finding is due to the fact that the reference period is stated in the introduction and not as part of the questions themselves, Igou et al., (2002) conducted a second experiment. In this study the reference period (either "today" or "the past six months") was part of each of the four questions. In three different conditions the anger question was placed either first, second, or fourth, thus, leading to a 2 (reference period) x 3 (question position) design. The result was that only when the anger question was asked first were there significant differences in the types of anger experiences reported. When the students were aware that the reference period was not unique to the target question, they seemed to not to pay attention to the reference period when interpreting the meaning of the question. The findings of these two studies are highly relevant for the SFSS and the current study. Like most mental health outcome measures, the reference period for the SFSS is stated for all 33 questions as part of the introduction and, thus, it is likely that the respondent will not pay attention to the reference period. However, it is important to consider several limitations of the studies cited above. First, the participants in all studies cited above were college students reporting about themselves. Thus, we do not know how much this generalizes to youths or to proxy reports by caregivers. In

addition, these researchers investigated only one type of event (e.g., experiencing anger). It could be that respondents would consider the reference period in situations of multiple questions if the target question refers to rare and salient events that have no or little ambiguity. An example may be a question about how often a youth was arrested by the police.

Retrieval

The knowledge about memory retrieval in the survey literature stems from the autobiographical memory research, which is an applied branch of memory research. Autobiographical memory research tries to understand the storage and recall of real, live events such as college graduation, weddings, and births of children. Although this branch of memory research has experienced an uprise in the last two decades the standard textbooks on memory are still dominated by laboratory studies. Nevertheless, different theories of autobiographical memory have emerged over the years. These theories differ mainly in the organization they impose on personal incidents in memory (Shum & Rips, 1999). Tulving (1983) proposed that episodes from your life might exist in memory in independent, minimally connected units. Kolodner (1984) sees life episodes as organized in hierarchies based on their distinctive properties. Other theories hold that life events are organized in thematically and chronologically structured histories or streams (Barselou, 1988; Conway, 1996). Shum and Rips (1999) distill as the central ideas of the theories, first, that autobiographical memory is memory for representation of personal events and, second, that people retrieve these event representations by describing a sufficient number of the event's parts and context, such as the location of the event, the people involved, the date and time.

Independently of these different theories, however, is that "by far the best-attested fact about autobiographical memory is that the longer the interval between the time of the event and the time of the interview, the less likely that a person will remember it." (Tourangeau et al., 2000, p. 82) According to Tourangeau and colleagues, this finding is not necessarily due to the passage of time by itself obliterating the event's details, but instead, "additional time makes it more likely that the person will experience similar events in the interim, and these later events interfere with the retrieval of the initial one." (p.82 ct) That is, events blur together making it easier to remember the overall pattern but more difficult to remember individual distinct events. In general, rates of forgetting are often found to follow a negative exponential function but that can differ based on the type of event (Tourangeau et al., 2000). It must be noted that many of the events to be recalled in these types of accuracy of recall studies are different from the types of events or experiences respondents of mental health outcome studies are asked to recall. The events studied include phone calls made (e.g., Belli, Schwarz, Singer & Talarico, 2000; Blair & Burton, 1987), making purchases (e.g., Blair & Burton, 1987), number of bank checks written (e.g., Blair & Burton, 1991), fishing and hunting (e.g., Chu, Eisenhower, Hay, et al., 1992), medical procedures (e.g., Loftus, Klinger, Smith & Fiedler, 1990), and types of food eaten (Smith, 1991). These are clearly different from recalling whether one got into trouble or felt worried. The studies by Winkielmann et al., (1998), Igou et al., (2002), and Thomas and Diener (1995) described earlier suggest that for these kinds of events or experiences the recall process may be more complicated and affected by other factors in the response process.

Aside from the temporal distance of the event to be recalled, there are several other characteristics of the event that can affect its subsequent accessibility for recall. The main ones are (a) the event's proximity to temporal boundaries, (b) its distinctiveness, (c) its emotional impact, and (d) for proxy reporters whether they were present during the event (Neter & Waksberg, 1964; Sudman and Bradburn, 1973; Tourangeau et al., 2000).

Proximity to temporal boundaries. As the reference period grows longer, it can become more difficult to map the time that the reference period refers to in one's own life. It is relatively easy to remember what happened in the past week and thus distinguish what happened two weeks ago from the events of this past week. With longer reference periods (e.g., six months) this task becomes more difficult and the exact boundaries of the recall period get blurry. One consequence can be either forward or backward telescoping, that is, either including events that occurred outside the specified reference period or excluding events that actually happened during the reference period. As a remedy survey researchers have found that anchoring the temporal boundaries to critical life events of the respondent (e.g., end of a semester) can improve accuracy of recall (Tourangeau et al., 2000). Such a strategy is difficult to implement in the context of youth outcome measures. Thus, if the goal is to improve accuracy in recall, a short reference period would be critical.

Distinctiveness. Distinct events, that is, events that are infrequent and atypical are more easily remembered than those that are frequent and typical (Tourangeau et al., 2000). Examples are of more distinct events are graduation from university and the loss of an important person while brushing teeth and eating a meal are examples of frequent and less distinct events.

Emotional impact. Another way events can be salient, and thus easier to remember, is when they are important to the person and when they have a high emotional impact (Tourangeau et al., 2000). This is interesting to consider in the context of the types of questions included in mental health outcome measures. "Getting into trouble," for example, can be a very salient event for a youth who typically behaves well and is mainly internalizing problems. Another youth who regularly acts up may not remember most of the times he gets into trouble because it happens so often. However, there is no indication whether the reference period would have a direct effect on this variable in the recall process.

Proxy reporters. Finally, Tourangeau et al., (2000) report that, not surprisingly, people who experience an event themselves are more likely to recall it than somebody who observed it or just heard about it. They cite research which confirms that proxy reports are often more likely the result of guesses or estimates than are self-reports.

Another important aspect to consider is the fact that some of the items in mental health outcome measures ask respondents to recall emotional experiences, such as: "How often did you feel worthless?" Recall of experienced emotions is different than that for events or behaviors. Robinson and Clore (2002) provide an excellent review of the relevant literature in this context and offer an accessibility model of emotional self-report. They propose that people access at least four types of knowledge when reporting on their emotions. First, people access their feeling directly (*experiential knowledge*). According to Robinson and Clore's review, this type of access is limited to reports of current events. Second, people attempt to recall the contextual details of emotional experiences (*episodic memory*). However, "the ability to recall contextual details", as

Robinson and Clore report, "declines quickly with passage of time." (p.935) More likely people will access their semantic memory when attempting to recall emotions. Semantic memory is not based on any particular event but rather consists of certain generalizations, that is, beliefs that are very stable. Robinson and Clore differentiate between *situation-specific beliefs* and *identity-related beliefs*. Most of us believe that completing a dissertation is associated with relief and happiness, which is an example of a situation-specific belief. Gender stereotypes in reports about emotions (e.g., "women are more emotional") is an example for identity-related beliefs. Of specific interest is the finding of several studies reviewed by Robinson and Clore that "one's past standing on a certain attribute is often inferred on the basis of one's current standing, in combination with beliefs about stability and change." (p.943) An illustrative example is a study by McFarland and Ross (1987). They tracked romantic relationships over a 2-month period and found that participants' retrospective reports at the end of the two months about their relationship two months earlier were systematically biased in the direction of the participants' current perception of the relationship. The participants in this case were influenced by their beliefs about the stability of their relationships.

This review by Robinson and Clore is relevant in regard to the expected differences based on the reference period. First, it is clear that when attempting to recall emotions over a three- or six-month period people will almost completely rely on their semantic memory. But, even for most cases in a two-week period people are likely to use their semantic memory when trying to recall emotions since access to experiential knowledge is not available and episodic memory of contextual events fades away rather quickly. Thus, in both cases one would expect that self-report on emotional experiences

is based on current beliefs about oneself or the other person in the context of proxy reports. Of relevance would be whether the person assumes stability or change when reporting on the frequencies of experienced emotions. If the person assumes stability of the emotion one would not expect real difference between the two-week version of the SFSS and the three- or six-month version. However, it is not clear what the person would do if they assume change in the emotional experience. If they believe that they feel better now than they used to three months ago, they may be tempted to select a different response if the longer reference period is used. Since the starting point of the reference period is always the current day and, thus, includes more recent times it may not be clear to the respondents how to respond. Should they select a lower rating because they feel better now or should they select a higher rating because they believe they felt worse three months ago. While I have no systematic empirical evidence for this that I could cite, I have heard anecdotal stories of interviewers who administer the CBCL and frequently encounter exactly this type of confusion within respondents. The reference period of the CBCL is "now or within the last 6 months." In summary, there may be only a very few cases in the comparison group who may have answered differently on emotion items than they would have had they been in the reference group because they believed they changed. Of course, this assumes that they pay attention to the reference period in the first place.

Aside from these characteristics of items and events, Bachman and O'Malley (1981) speculate that "one of the reasons why events in the past are underreported in surveys may be that many individuals are simply not sufficiently careful or motivated in their role of respondent." (p. 546) While I have no systematic empirical information

about the youths' and caregivers' motivation to complete the SFSS available, in conversations with clinicians who participated in a pilot study, I learned that many youths as well as caregivers were resistant to completing the questionnaire. Another indicator may be that the respondents in the psychometric study had very little time to answer each question. Tourangeau et al., (2000) note that with less time per question the accuracy of recall declines and vice versa. On average the respondents had approximately 8 minutes for answering the 33 questions of the SFSS. That is, they had approximately 24 seconds per question. Clearly, that is not very much time to recall individual events. This makes it likely that the respondents do not actually attempt to recall individual events for each question of the SFSS but use an estimation strategy instead. This will be discussed as part of the next section.

Judgment

Many survey questions require the respondents not only to recall events from memory but also to combine, summarize, or estimate the information in some way. In certain cases the respondent may even skip the recall process all together and proceed directly with estimation. This judgment process is especially relevant for questions about the frequency of events like the questions of the often mentioned outcome questionnaires. Based on experimental memory research, decision theories, and studies of frequency estimation in surveys Tourangeau et al., (2000) describe several broad groups of strategies respondents use for answering frequency questions.

The first one, recall of specific information, includes three different ways of remembering events and then the counting by the respondent. First, respondents simply count all relevant events they can remember. Second, respondents try to remember events

by certain domains and then add the number of events in each domain together to obtain a total count. When asked about arguing with adults, for example, a youth may first think of all arguments he had with his parents, then of all the arguments with his teachers, and finally with all other adults he interacted with during the reference period. Third, a strategy, which Tourangeau et al., termed recall-and-extrapolate, involves recalling several episodes and using them to estimate a rate of occurrence which is then applied to the total reference period. A youth may recall three times during the last week where she was extremely tired. When asked about how often she was unusually tired during the last three months, she simply infers that is must have been three times a week for the duration of those three months even though she does not really remember how often it was in each week except for the first one. It seems clear that if this strategy is used the differences in reference periods, especially longer ones are reduced.

The second group of strategies encompasses estimations based on generic information. Because memories of similar and regular events can blend together as one generic representation of this group of events, it can be difficult to recall specific individual events. In cases like these "respondents may resort to recalling generic information instead, information such as the typical rate of occurrence for the behavior in question." (Tourangeau et al., 2000, p. 148). It appears logical that respondents would more likely opt for this kind of strategy if the reference period is longer and the events to recall occurred further back in time. For example, I will report the findings of a series of studies of dietary recall by Smith (1991). Two groups of respondents' kept a daily food diary for two or four weeks. Recall measures were completed at the end of the two or four week diary keeping. In addition, recall measures were administered to both groups

several times in the weeks following the initial recall measure. Besides the generally low correspondence of the recall measures and diary in all conditions, it is interesting while the correspondence rate between diary and recall dropped quickly in the weeks following the initial recall measure, after about six weeks the correspondence rate leveled off and remained at about 30%. Smith concluded that in these cases the respondents probably listed food items that they routinely eat rather than to actually recalling the exact items they ate.

If the number of events in question is very important, if it is asked about often enough, or if it is small enough, the respondent may simply be able to recall an exact tally. As evidence for this third group of strategies, Tourangeau et al., reference a study by Brown and Sinclair (1997) in which a significant number of respondents who reported having at least eight sexual partners claimed that they used this strategy of recalling an exact tally. It is not necessarily clear whether the reference period is likely to have an influence on this type of strategy. The same is true for the fourth group, which Tourangeau et al., (2000) refer to as estimation based on a general impression. This residual category is for respondents who report using no particular process and includes things like guessing or rough approximation.

Several factors such as the number, the distinctiveness, and the regularity of events have been found to influence the selection of the response strategy (Tourangeau et al., 2000). Among these is the reference period. Blair and Burton and others (Blair & Burton, 1987; Burton and Blair, 1991; Chu et al., 1992) found that as the reference period gets longer respondents are less likely to use the strategy where they recall each event and count the events to get the total number. In the study by Blair and Burton (1987) on

recalling telephone calls, they found that for a reference period of two weeks more than half (56%) of the respondents used an enumeration process, while for a two-month period only a quarter of the respondents used enumeration. Almost no (4%) respondents used an enumeration process for a six-month reference period. The findings by Burton and Blair (1991; recall of checks written) and Chu et al., (1992;) are similar.

Response

Tourangeau et al., (2000) discuss two aspects of the process of selecting the final response: the mapping of judgments to survey answers and the editing of responses. As mentioned previously, outcome measures usually use rating scales with two or more answer categories. There are a variety of processes respondents seem to utilize when mapping their answer to the prescribed ordered answer categories. The main theory Tourangeau et al., refer to is the range-frequency model by Allen Parducci (1965, 1974). This assumes that respondents have the tendency to begin by mapping the most extreme stimuli to the scale endpoints and then proceed by using this as anchors for mapping the remaining items. The frequency principle of this model holds that respondents tend to distribute their answers evenly across the different answer prompts. According to this model a youth, for example, may begin answering an outcome questionnaire by selecting "never" for being arrested by the police. Then, he would select "very often" for "getting into trouble" followed by "sometimes" for "getting into fights". Towards the end he realizes that he has not selected "often" once. Thus, he selects "often" for "argue with adults." While there is some empirical support for this general model there are often additional processes present. One of the biases respondents tend to have is a positivity bias when rating others, they tend to navigate towards the positive end of the scale, avoid

negative numbers, and use the extreme ends of scales only sparsely. In regard to the reference period it could be that youth have different ideas of what "very often" means theoretically for a two-week period compared to a six-month period. This seems to be supported by the difference in the interpretation of the term "anger" when different reference periods were used as reported in the study by Winkielmann et al., (1998) cited earlier in this chapter.

Another very likely scenario is that the youth will edit their responses. It is a well-known phenomenon in the survey literature that respondents edit their responses if they consider the topic of the question as too sensitive. This is especially prevalent among teens who often fear getting into trouble if they disclose a socially undesirable or even illegal behavior. Tourangeau et al., (2000) cite a range of studies that demonstrated that teens tend to underreport socially undesirable behaviors such as smoking and illicit drug use. There seems to be less underreporting if the survey is self-administered. Many outcome measures for youth contain questions about smoking, illegal drug use, as well as other undesirable behaviors such as having contact with the police, being in fights, hanging out with people who get into trouble, etc. In addition, the youth must assume that either the clinician or somebody else in the office of the clinician will see their answers. In some cases when the results of these measures are provided as feedback it is even explicit that the clinician will see the answers even if it is in summary format. Thus, we must assume that youth underreport many of the behaviors they are asked about. This is confirmed by the finding that the severity scores of the youth are often much lower than those by others such as the caregiver or the clinician, especially for externalizing behaviors (e.g., Achenbach et al., 1987). There is some evidence that would suggest that

the reference period could make a difference in this editing process. In these studies respondents were more likely to underreport recent drug use than drugs used in the distant past (Tourangeau et al., 2000). Thus, if the reference period is long (e.g., six month) the youths may feel safer to admit that they sometimes used drugs (or any of the other undesirable behaviors) because they can claim that it happened a while ago but is currently not a problem.

Application of Theory and Empirical Evidence to the Current Study

A somewhat simplified model of the complex response process discussed above as it applies to the response of youths to the SFSS is presented in Figure 1. This model helps to determine whether, on average, one would expect to find differences between the reference group and the comparison group or not. The thickness of the arrows represents the expected likelihood that a person will follow this specific path in the response process. The stimulus is the same for all respondents (i.e., a 33-item questionnaire with vague answer choices and the reference period stated in the instructions) except for the difference in the reference period. During the *comprehension* stage the key question is whether the youth even pays attention to the reference period. Only if they do, will this difference in the stimulus be able to affect the outcomes. The literature discussed earlier suggests that, because of the number of items that all have the same reference period that is stated at the very beginning as part of the instructions, it is very likely that the youths will not pay attention to the reference period stated in the instructions and, consequently, will not differ significantly in their answer choices.

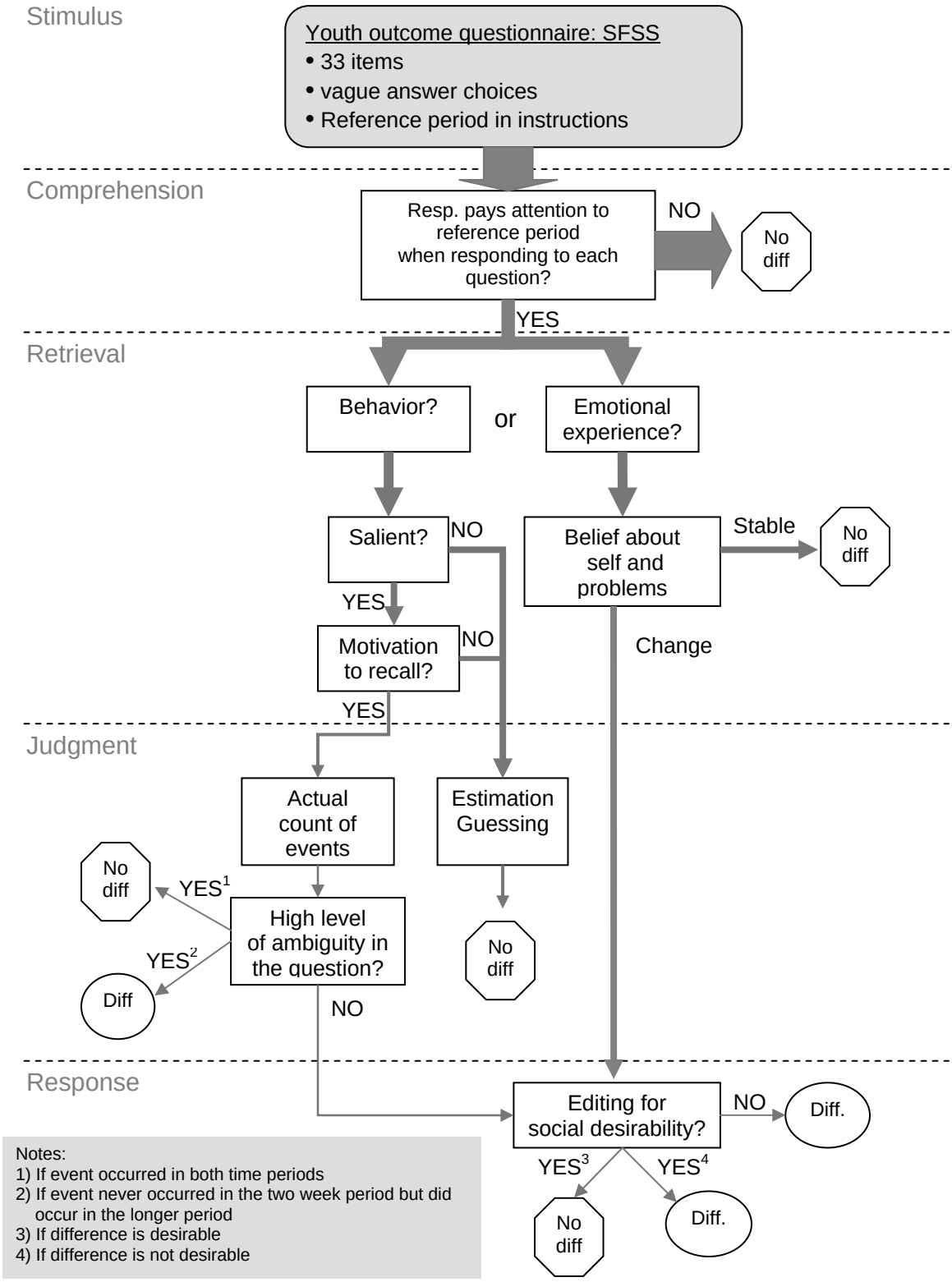

Figure 1: Model of Response Process Applied to Current Study

For those youths who actually pay attention to the reference period, the *retrieval* process will likely differ depending on the content of the item. If the item asks the youth to recall behavioral events, such as getting into trouble, it will matter how salient the events were to the youth. Salient events are easier to recall than less salient ones. Thus, if the events are not very salient, the youths are likely to estimate or guess based on their current perception or their typical behavior. It is not likely that they will use a different estimate for the last three or six months than they would for the last two weeks. In fact, the best ground for estimating the typical rate for the longer reference period is what it has been during the most recent weeks. If, however, the events are salient, the youth could actually try to recall these events. This is most likely for rather infrequent types of events. Whether the youth actually tries to recall the events, however, will depend on their motivation. If they are not very motivated, they will likely not make the effort to accurately recall and count the events but will provide an estimate or guess instead. As argued earlier, it is very likely that the youths are not very motivated to spend a lot of effort in accurately answering these questions, especially given the short amount of time to do so.

If the item is asking the youth to report about the frequency of experienced emotions, one would, based on the review by Robinson and Clore (2002), expect that they will generate their answers based on their beliefs about themselves and their problems for both types of reference periods. If the youths also believe that they have been stable for a while, the difference in the reference period is likely not to be meaningful. On the other hand, if they believe they have significantly changed, they may be tempted to answer differently if the three- or six-month reference period is presented

to them. It would depend on how they interpret the meaning of the three- or six-month period. That is, if they believe they experience the type of events (e.g., worrying a lot) now much less than they used to three months ago, they may interpret the three months period as referring to that time three months ago and select a higher frequency choice (e.g., "*often*") than what they would have endorsed if the reference period was two weeks (e.g., "*sometimes*" or "*hardly ever*").

During the *judgment* stage the expectation of differences depends on the type of strategy the youth is likely to employ. If they simply estimate or guess, it is likely that their estimate for the longer time period is inferred from the experiences during the more recent weeks as discussed earlier. However, if they actually recalled the events and enumerated them, then they will have to match the actual number to the vague quantifiers. The studies by Winkielmann et al., (1998) suggest that this process differs depending on how vague the question is and how clearly the key terms (such as "anger") are defined. If the question is vague, differences in actual numbers of recalled events in both conditions (2 weeks and 2 or 3 months) could lead to the same answer because the answer choice is relative to the reference period rather than an absolute judgment. However, if the youth is very certain that the event never happened (e.g., they do not recall ever drinking any alcohol), it is likely that they select ("never") in any condition. If they are not certain, they are likely not to select "never" unless they edit their responses for other reasons, such as social desirability. Editing for social desirability can lead to either differences based on reference period used or no differences. This will depend on whether the social desirability is affected by the length of the time period one is referring to and the perceived consequences of true disclosure.

For the caregiver I would expect a similar process. However, as proxy reporters, different types of events may be more or less salient to them and the semantic memory that is relevant for emotional experiences are beliefs about the youths not themselves. It is also important to note that the purpose of the current study is not to test this model. This model simply served to guide the formulation of the hypotheses. However, the model highlights some points of interest for future research as I will discuss in the final chapter.

To summarize, the review of the existing literature and the application of the major findings to the current context lead me to believe that using different reference periods will not have a significant impact on how people respond to the SFSS. That is, on average, I would not expect to find meaningful differences between the groups. But, what exactly do I mean by meaningful differences? This will be discussed next.

Hypotheses

In the research as well as the clinical practice world a difference becomes meaningful if one would draw different conclusions or make different decisions based on that difference. For example, if two randomized groups who have received different treatments differ on their total score on the outcome measure by the end of the study we assume that the treatments work differently. (Of course, this depend on the exact design which is often more complicated than this example). Now assume that in another study with the same design we use a measure with a different reference period and it would change the way the youths or caregivers respond to the questions significantly and, thus, potentially change the total scores. It may be that because of these changes in the scores, the difference between the two groups becomes insignificant and we would, thus,

conclude that the two treatments are not different. In another scenario a client is referred to a provider and they administer a six-month version of the SFSS at intake. The client is in the high severity range. Two weeks later they administer the two-week version of the SFSS and find that the client has moved significantly out of the high-severity range. Could this be simply due to the different in the reference period or has the client really changed that much in two weeks of treatment? While the relevance of the reference period should be investigated for both types of cases, that is, the group level and the individual level, in this study I will primarily focus on differences that affect the outcomes at the group level, such as averages and proportions. While both types of situations are equally important, the dataset that was available to me was best suited to investigate differences at the group level due to the original cross-sectional experimental design.

What aspects of a measurement scales are important in the sense that they could affect inferences at the group level? In the psychometric development of measurement scales one is often concerned with four aspects: central tendency and variability, validity, and reliability. All four of these scale characteristics can affect the inferences that are made based on scores from a scale score. If a change in reference period would affect any of these characteristics, it would be a meaningful difference. Thus, using the available dataset I will test if there are no meaningful differences in regard to those four aspects.

Hypothesis 1—Central tendency

The most common measure of central tendency is the mean score.

Hypothesis 1a: The **difference of the mean** SFSS score (as rated by the youth) between the reference and the comparison group is not meaningfully different.

Hypothesis 1b: The **difference of the mean** SFSS score (as rated by the caregiver) between the reference and the comparison group is not meaningfully different.

How a meaningful difference is defined exactly and how I will determine if there is no difference, on average, will be discussed in detail in the analysis section.

Hypothesis 2—Membership in high-severity range

There are several ways to measure variability among participants in a sample. The most commonly used one is the variance or standard deviation. However, a more interesting and useful aspect of the variability in a sample in the context of clinical scales is how many clients fall into the clinical or high-severity range according to the cut-off scores established for a specific scale. This goes back to a widely-accepted understanding of clinically significant change in the context of psychotherapy outcome studies that was proposed by Jacobson and colleagues (Jacobson, Roberts, Berns, & McGlinchey, 1999; Jacobson & Truax, 1991). According to these researchers a change is considered clinically significant if the following two criteria are met: "(a) The magnitude has to be statistically reliable and (b) by the end of therapy, clients have to end up in a range that renders them indistinguishable from well-functioning people." (Jacobson, et al., 1999, p. 300) Thus, if the use of different reference periods does not lead to a meaningful difference in the proportion of youths who would be classified as "clinical" or "high

severity" than I have demonstrated another important aspect of equivalence between the reference and the comparison group.

Hypothesis 2a: The proportion of youths who are **in the high severity range** (based on the youths' ratings) in the comparison group is not meaningfully different of that in the reference group.

Hypothesis 2b: The proportion of youths who are **in the high severity range** (based on the caregivers' ratings) in the comparison group is not meaningfully different of that in the reference group.

Again, the exact definition of a meaningful difference will be discussed later.

Hypothesis 3—Validity

The third aspect of the SFSS to be investigated is the validity. Several different types of validity have been proposed (see Shadish, Cook, & Campbell, 2001 for a review). One that is commonly used in the validation of psychological scales is concurrent validity (Meyer et al., 2001). That is, the validity of the scores from a newly developed measure is demonstrated by the fact that it correlates highly with other similar and well established measures. In the case of the SFSS, for example, we correlated it with the CBCL/YSR, the YOQ, the SDQ, and the CAFAS. Should one expect that the correlations between measures with reference periods that are matched or close to each other (e.g., six months and six months) are higher than those between measures with very different reference periods (e.g., six months and two weeks)? Based on the deliberation earlier, this is probably not the case, but that remains to be shown. This leads to the third hypothesis.

Hypothesis 3a: The **correlation** of the SFSS scores (rated by the youth) with the YSR in the comparison group is not meaningfully different from the respective correlation in the reference group.

Hypothesis 3b: The **correlation** of the youth SFSS scores (rated by the caregiver) with the CBCL in the comparison group is not meaningful different from the respective correlation in the reference group.

I selected the YSR and the CBCL as the test case because (a) they represent the most well-established and widely used outcome measures for youth, (b) the group that completed these measures is the largest relative to the other groups, and (c) the difference between the reference period of the SFSS and the CBCL/YSR is the greatest (2 weeks compared to six months). Thus, if there is no difference in this case, it reasonable to assume that they should also be no difference for measures that are closer in their reference periods.

Hypothesis 4—Internal reliability

Finally, it is important to establish the reliability of a scale. That is, does the scale measure a trait or attribute consistently over repeated administrations (Shadish, et al., 2001). If a scale is unreliable, for example, it will affect the estimate of the magnitude of the correlation of the construct measured with this scale with any other construct assessed by a different measure (Cohen, Cohen, West, & Aiken, 2003). The two major types of reliability are internal reliability and test-retest reliability (Shadish, et al., 2001). Because the latter requires measurement at two different time points, which was not available in the current dataset, I will focus on the first one in formulating the fourth and final hypothesis:

Hypothesis 4a: The estimate of the **Internal Reliability** of the SFSS scores (rated by the youth) assessed in the comparison sample is not meaningfully different from the respective reliability estimate in the reference group.

Hypothesis 4b: The estimate of the **Internal Reliability** of the SFSS scores (rated by the caregiver) assessed in the comparison sample is not meaningfully different from the respective reliability estimate in the reference group.

These tests for differences are not exhaustive. There are other interesting aspects to be investigated, such as the sensitivity of each version to change over time or differences in individual items. However, based on the data available to me and the scope of this dissertation, I decided to limit the current investigation to these four tests.

CHAPTER II

METHODS

The purpose of this chapter is to provide the reader with important information about the study. I will begin by describing the instruments used for this study followed by the procedures that were used in generating the original dataset from which I extracted the current dataset for the secondary analysis. This includes a description of the randomization process. I will then provide a description of the sample that was used for the current study.

Instruments

Symptom and Functioning Severity Scale (SFSS)

The SFSS was developed to monitor the development of a youth's mental health status over the course of mental health treatment. It is not meant as a diagnostic assessment instrument, but rather as a brief measure (33 items) for frequent administration to monitor youth mental health outcomes. The SFSS covers a range of common internalizing and externalizing problems in youth with items based on four of the most prevalent childhood disorders: attention deficit hyperactivity disorder (ADHD), conduct/oppositional disorder, depression, and anxiety. In addition, it includes items related to peer and family relationship problems. Respondents are asked over the last two

weeks (or any other reference period that was tested), how often the youth experienced the behavior, emotion, or cognition described by each item. The answer options are *Never, Hardly Ever, Sometimes, Often* and *Very Often*. Example items are "In the last two weeks, how often did you get in trouble?" and "In the last two weeks, how often did you feel worthless?"

There are three versions of the SFSS: one for youths 11-18 years old, one for caregivers, and one for clinicians (only the former two were used in this present study). For each of these versions a total score is calculated providing a general indicator of the youth's mental status from the perspective of each respondent.

The SFSS was developed over the course of several years by Leonard Bickman and his team and has been tested in several pilot studies and revised based on the pilot study findings. It has been evaluated with cognitive interviewing techniques. The current version was tested for its psychometric properties in the psychometric study from which the data for this present study were drawn (Bickman, et al., 2006). I will demonstrate the quality of the SFSS as a reliable measurement scale with the current data below using a Rasch measurement approach.

Child Behavior Checklist (CBCL)

The Child Behavior Checklist (CBCL; Achenbach, 1991a) is a popular scale used to assess children's emotional and behavioral problems. It is a parent-report checklist of 118 behavioral and emotional problems (e.g., "cruel to animals," "sad, unhappy, or depressed"). For each item, parents report whether their child has the problem by circling 0 ("Not True"), 1 ("Somewhat or Sometimes True"), or 2 ("Very True or Often True"). The reference period is stated as "now or in the last six months." Several scores and sub-

scales are available. For the purpose of testing hypothesis 3 I will use the total problem score.

Youth Self-Report (YSR)

The YSR (Achenbach, 1991b) is a youth self-report version of the CBCL with 112 items and can be completed by youths with 5th grade reading skills, or it can be administered orally. Its competence and problem items generally parallel items 6-18 of the CBCL. The YSR includes questions allowing open-ended responses for items covering physical problems, concerns, and strengths. Youths rate themselves for how true each item is now or was within the past six months using the same three-point response scale as for the corresponding version of the CBCL--0 ("Not True"), 1 ("Somewhat or Sometimes True"), or 2 ("Very True or Often True"). In addition, the YSR has 14 socially desirable items that most youths endorse about themselves. The YSR scoring profile provides raw scores, *T* scores, and percentiles for several types of scales and sub-scales. Scales are based on 2,581 high-scoring youths and normed on 1,057 nonreferred youths. For the purpose of the present study I will use the total problem score.

Procedures

The original study was conducted from June-September, 2005 in order to obtain descriptive information about youths served by a large national for-profit mental health service provider, to test the psychometric properties of measures in the Peabody Treatment Progress Battery (PTPB), and to obtain feedback from clients, adult caregivers, and clinicians on their perceived utility of the measures. Twenty-eight offices owned or managed by the mental health company, located primarily in the Eastern and

Midwestern U.S., participated in the study. A research team from Vanderbilt University led by Leonard Bickman provided consulting in the planning of the study and was also responsible for data processing and the analyses as described below.

Eligibility Criteria

Eligible respondents were youth ages 11 to 18, their primary caregiver, and their primary clinician. All clients in the appropriate age range who had receive individual treatment as part of mental health services at any of the participating offices during the duration of the data collection of four weeks were eligible as long as they had received at least one week of treatment. Clients were encouraged to participate in the psychometric study only if the clinician thought the client was able to understand questions in the SFSS and other instruments in the measurement battery. If clients were not able to comprehend either the English or Spanish language, they were not eligible to participate.

Randomization

The research question regarding the relevance of the reference period that I am investigating in this dissertation was anticipated in the planning stages of the original study and a series of randomized experiments was embedded in the main study as a validity test. Of the 28 participating regional offices, 26 participated in this randomized experiment while two offices participated in a separate test-retest reliability study. The 26 offices were divided into five groups based on the number of clients they served and the counselors' familiarity, if any, with administering any of the validity measures. This helped assure an adequate number of cases in each test group as well as eased the burden of completing new forms. Participants in each office all received the same validity

measure. Half of the respondents in each office were given the SFSS with a 2-week reference period. The other half received the SFSS tailored to have the same reference period as the validity measure administered at that site (see Table 2 for SFSS and validity measure reference periods and group validity measure assignments).

Prior to shipping materials, the envelopes were interleaved so that every other envelope included the same pairing. This procedure helped assure a balanced number of the two pairing versions for each office. Regional directors, who were responsible for distributing the envelopes within their office, were blind to this procedure. Table 2 provides an overview of the grouping.

Table 2: Original Study Groups

Group	Reference period for SFSS	Reference period for Validity Measure	Validity Measure	Number of Offices
1a	2 weeks	1 week	YOQ	2
1b	1 week	1 week	YOQ	
2a	2 weeks	1 month	SDQ	5
2b	1 month	1 month	SDQ	
3a*	2 weeks	3 months	CAFAS	5
3b*	3 months	3 months	CAFAS	
4a*	2 weeks	6 months	SDQ	8
4b*	6 months	6 months	SDQ	
5a*	2 weeks	6 months	CBCL/YSR	6
5b*	6 months	6 months	CBCL/YSR	
Total				26

*Groups included in the current study.

The purpose of the current study is to compare the two-week reference period version of the SFSS to versions of the SFSS with longer reference periods matched to the reference period of the validity measure. With reference periods of three and six months

groups 3, 4, and 5 all qualify to be included in the current study. The contrast between the one month version of group 2 and the two week version is not big enough for it to be included for the current study. The comparison in group 1 was a one-week version and, thus, clearly did not qualify. To simplify the analyses and to increase statistical power I created two samples from these three groups: (1) the *reference* sample that included all youths and caregivers who completed the two-week version of the SFSS within groups 3, 4, and 5 and (2) the *comparison* group that included all youths and caregivers who completed the three- or six-month version of the SFSS[1]. These samples were used for testing hypotheses 1, 2, and 4. Since hypothesis 3 includes the use of the CBCL and YSR, only the respondents in group 5 could be used to test that hypothesis.

Data Collection

The youth, a primary adult caregiver, if present, and the clinician completed the measures at the end of a session[2]. In order to minimize burden, the full set of measures used in the main study were divided into two booklets of different (not "repeated") measures and administered at two consecutive sessions (on average, one week apart). The SFSS and the validity measure (e.g., the YSR and the CBCL) were placed in the 1st booklet. The SFSS was the first measure in the 1st booklet only preceded by some background questions. For the youth and the caregiver the SFSS was followed by two other short scales (a life satisfaction scale and a hopefulness scale for the youth and a life satisfaction and caregiver strain measure for the caregiver) and the validity measure,

[1] The rationale to combine the 3 and 6 months samples was to obtain sufficient statistical power for the hypothesis testing. I felt justified to do this based on theoretical considerations as well as on an explorative analysis of the similarities between the two samples.
[2] The rationale for having the respondents complete the measures at the end of the session was that for the actual use of the measures in the future it is also planned to have them completed at the end of a session.

which came last. The clinicians completed the SFSS, the validity scale, and several questions about themselves, the client, and the survey. The second booklet included measures on common factors (e.g., therapeutic alliance and treatment motivation) and was administered in the next available session.

Clinicians were allowed to read questions to youths and adult caregivers from a reading copy if necessary, but were instructed not to help with answers. All youth and adult caregiver measures were available in English and Spanish[3]. Offices were asked to administer both booklets to all eligible clients within four weeks then ship their completed materials to Vanderbilt. This time frame was optimistic; data were received from the 1st region about seven weeks after the study's start date.

Confidentiality

All data were obtained by the mental health company as part of its continuous quality improvement (CQI) initiative – Contextualized Feedback Intervention and Training (CFIT). The researchers from Vanderbilt had no contact with participants either for recruitment or data collection. Names or other information that could identify respondents were not sought or obtained. To link data collected about the same youth from multiple respondents at two time points, the researchers developed a unique identification (ID) number for each client that was based on a concatenation of a unique region code (e.g., 01-28), the last four digits of the clinician's social security number, and a unique number (001-199) that the researchers assigned consecutively to each youth within each region. The region and youth IDs were preprinted on all forms and

[3] All Spanish measures have been translated and back-translated.

envelopes; clinicians recorded their ID on all forms/envelopes upon their receipt. All forms for each youth were enclosed in individual envelopes prior to shipping to offices. In order to facilitate data collection and help clinicians keep materials for the same client together, peel-off labels on the envelopes were used. Clinicians were instructed to write their name and the client's name on the label, but to remove the label before they shipped materials back to Vanderbilt. All data received and maintained by Vanderbilt included only this unique non-sensitive participant ID.

Data Processing

A multi-step process began once data were received. First, the number of booklets received was logged by region, respondent type (youth, adult caregiver, clinician) and type of booklet (1^{st} or 2^{nd}). Second, a detailed protocol was used to check for data quality, including problems of respondents recording two answers for the same item or highly suspect response patterns that would suggest invalid data. Coin toss was used to determine which among two answers for the same item to code—if the responses represented adjacent categories and were not contradictory. If agree and disagree were both endorsed, the item was coded as missing. Data that remained ambiguous were considered missing.

Unusual response patterns were reviewed independently by two raters. There were remarkably few instances of unusual response patterns, and the raters nearly always agreed when one presented itself. The project's data manager made the final determination in the event of inter-rater disagreement. Cases with any response pattern were flagged by measure, so that they could be excluded as needed during analysis. Safeguards used for data entry were initial cleaning, and data were entered into an

ACCESS® database once, then again. Special programming alerted data entry staff to discrepancies between the two entries as well as entry of out-of-range values for each variable. The ACCESS database was translated into SAS system files. Univariate statistics (e.g., frequencies; means) for each variable were generated and examined for accuracy and corrections were made where indicated.

Completion Rates

There was no reliable information available from the offices in regard to how many eligible clients were served during the period of the data collection. Thus, I am not able to assess the degree of representativeness of the sample in this study. There was some reasonable suspicion that some of the most severe cases were not included because of the crisis situation in their home and the counselor's inability to collect the data. However, further investigations into this possibility could not confirm this suspicion.

Data Extraction for Secondary Analysis

For the purpose of this secondary analysis I extracted a subset of the original data file that contained only the variables for each of the measures I am using for the analyses (SFSS and CBCL/YSR) as well as for some background information about the respondents such as age, gender, and length of current treatment. The use of this subset of data for the purpose of this secondary analysis was approved by the Vanderbilt IRB. For reasons explained above only data from groups 3, 4, and 5 were included.

Sample Description

Youths

Of the 431 youths in the current sample[4], 42% reported to be female. Thirty-five percent were African American and 60% were Caucasian. Nineteen percent were Hispanic. However, only 3% of the sample reported Spanish as their primary language while almost all 97% consider English to be their primary language. On average, youths were 15 years old, with 20% between 11-12 years of age; the majority (64%) were between 13-16, and 14% were between 17-18. Many of the youth were referred to the health care provider for mental health services by the judicial system which explains why over half of the sample (57%) have been arrested at some point in their life and 35% have been convicted for their crimes according to the youths' self-report. On average, youth had been in treatment about nine months (262 days) at the time they completed the 1st booklet. Eight percent had received services for less than a month, while 24% had been in treatment between 1-3 months, 19% 3-6 months, 14% 6-12 months, and 14% over a year. Based on caregiver reports, over half (53%) of youths had been diagnosed with a mental health disorder at some time before or during current treatment. See Table 3 for more youth sample data.

[4] Not all youths and caregivers provided complete background information. The numbers and percentages are based on the current sample of 431 youths and 325 caregivers. Thus, any discrepancies to 100% represent those respondents with missing information. The caregiver sample is smaller because in several cases the caregivers were not involved in the treatment or were simply not present during the session data were collected. The numbers above also exclude 11 youths (2%) who had more than 15% of the items on the youth SFSS missing and 8 caregivers (2%) who had more than 15% of the items on the caregiver SFSS missing. These were excluded because they were not included in any of the analyses due to their high level of missing items on the core measure that is used in all analyses.

Table 3: Sample Descriptors

	Whole Sample				By Experimental Group	
	M	**(SD)**	**%**	**(N)**	**2 Weeks**	**3/6 Months**
YOUTH SAMPLE					**n = 200**	**n = 231**
TOTAL			**100%**	**(431)**	**100%**	**100%**
Female			42%	(179)	39%	44%
Male			58%	(248)	61%	55%
Age	14.8	(1.93)				
Age 11-12			20%	(86)	23%	17%
Age 13-16			64%	(274)	63%	64%
Age 17-18			14%	(60)	13%	15%
Caucasian			60%	(260)	63%	58%
African American			35%	(149)	34%	35%
Length in Tx (in days)	262	(385.71)				
0-1 month in Tx			8%	(33)	7%	8%
1-3 months in Tx			24%	(104)	23%	25%
3-6 months in Tx			19%	(84)	22%	18%
6-12 months in Tx			14%	(60)	14%	14%
> 12 months in Tx			14%	(62)	15%	14%
Ever arrested			57%	(244)	57%	57%
Ever convicted			35%	(153)	36%	35%
Ever diagnosed			53%	(230)	50%	56%
CAREGIVER SAMPLE					**n = 146**	**n = 179**
TOTAL			**100%**	**(325)**	**100%**	**100%**
Female			87%	(282)	80%	92%
Male			12%	38)	14%	10%
Age	44.9	(10.76)				
Age 18-30			4%	(14)	3%	6%
Age 31-40			32%	(105)	25%	38%
Age 41-50			30%	(97)	31%	29%
Age > 50			27%	(88)	32%	23%
African American			38%	(122)	38%	37%
Caucasian			59%	(193)	57%	61%
Birth parent			42%	(138)	38%	46%
Family member			14%	(46)	14%	14%
Foster parent			37%	(120)	38%	36%
Known youth < 1 year			17%	(56)	19%	16%
No High School Degree			16%	(53)	16%	17%
Bachelor Degree or Higher			14%	(45)	12%	15%
Never married			12%	(38)	8%	14%

Married / Living as married	46%	(150)	49%	43%
Widowed	7%	(22)	4%	9%
Divorced / Separated	30%	(98)	29%	31%
< $22,000	38%	(124)	36%	40%
$22,000 - 30,999	18%	(59)	17%	19%
> $31,000	30	(98)	31%	29%
Ever diagnosed	20	(64)	20%	19%

Caregivers

On average, caregivers were 45 years old. Of the 325 caregivers in the sample, only a few (4%) were younger than 30 years old. About a quarter (27%) were over 50 years old with 85 being the oldest reported age. The majority (87%) were female. About 38% were African American and 59% Caucasian; Eleven percent were of Hispanic descent. About 16% of caregivers did not attain a high school diploma; 14% earned a bachelor's or higher degree. A little bit more than a third (38%) had annual household incomes less than $22,000; 30% had incomes of $31,000 or more. About a third of caregivers were divorced or separated, about half were currently married or living as married, 12% were never married, and 7% were widowed. About a fifth of the sample (20%) reported ever being diagnosed with an emotional, behavioral, or substance use problem.

Regarding their relationship to the youth, nearly all (97%) caregivers considered themselves the youth's primary caregiver. Less than half were the youth's birth parent (42%); a comparable number were foster parents (37%), with grandparents or other family members making up the balance. A significant number of caregivers (17%) knew the youth for less than a year. Most of those reported they knew the youth at least fairly

well, with only 14 caregivers reporting they did not know the youth very well. Overall, both samples are well balanced between the reference and comparison group suggesting that the randomization worked successfully.

Pre-Analyses

Examining the psychometric properties of the SFSS and creating measurement scores

Almost without exception youth outcomes measures use Likert scale ratings such as Never, Sometimes, and Often. In the data processing step each of these ratings is assigned a numerical code. While it can be correctly assumed that these categorical ratings are ordered, they do not qualify directly as a measure since "the very idea of measurement implies a linear continuum of some sort, such as length, price, volume, weight, age. When the idea of measurement is applied to scholastic achievement, for example, it is necessary to force the qualitative variable into a scholastic linear scale of some kind." (Thurstone, 1959 (1928), p.218). Simply adding each numerical code to a total score or taking the average does not do the job. In fact, if following arithmetic rules, it is not permissible to sum categorical data. While for a long time this fact has been ignored in the social science field, "new rules of measurement" (Embretson, 1996, p.341) have now been defined. The essence of these new rules is to use measurement models to create linear interval-level scales.

The development of these new rules of measurement has been facilitated by the introduction of Lord and Novick's (1968) now classic book on model-based measurement as well as Rasch's (1960) influential book on probabilistic models. While

both of these theories of testing have many things in common and are often generally referred to as Item Response Theory, there are many followers of these approaches who would consider them incompatible paradigms (see Andrich, 2004). Because it would go beyond the scope of this paper to follow this debate, I will limit myself by stating that for the purpose of this inquiry I am using a Rasch measurement model. Part of my rationale for this decision is that the purpose of using it is to create measurement scores for further analyses, not to explain the data with the best fitting model. Other factors in favor of the Rasch model are its simplicity, the availability of affordable and well-functioning software. Another advantage of using the Rasch model is that raw-person and item scores are minimally sufficient statistics for person and item parameters (Wright & Masters, 1982) just like the sample mean is a sufficient statistic for estimating the population mean given a large enough sample.

Of the family of Rasch measurement models the appropriate one for creating measurement scores from the SFSS raw scores is the rating scale model (Andrich, 1978; Wright and Masters, 1982; Wright and Mok, 2004). Besides the person ability (*Bn*) and the item difficulty score (*Di*) the rating scale model also includes a difficulty estimate (*Fx*) for the item threshold. The item threshold is the person measure score expressed in logit units for which a person with that score is equally likely to endorse one answer category over another (e.g., *Hardly Ever* over *Never*). The general Rasch rating scale model is described by the following equation:

$$\ln\left(\frac{P_{nix}}{P_{nix-1}}\right) = B_n - D_i - F_x \qquad (1)$$

That is, the natural logarithms of the odds ratio of the probability *P* of person *n* choosing category *x* of item *i* over selecting the previous category is modeled as a difference function of the person ability (*Bn)*, the item difficulty (*Di*), and each threshold estimate (*Fx*). This additive functionality plus the fact that the person, the item, and the threshold raw scores are sufficient statistics for estimating the respective model parameters are advantages of the Rasch rating scale model.

In conducting the analysis and reporting the results I am following the guidelines established for the *Journal of Applied Measurement* (Smith, Linacre, and Smith, 2003), the main scholarly outlet for Rasch measurement articles. For creating measurement scores and evaluating the fit of the data to the Rasch model I am using Winsteps® (Version 3.61.1), the most commonly used Rasch measurement software developed by Linacre and Wright. Winsteps® uses Joint Maximum Likelihood Estimation which is explained in Linacre (2004).

Youth Version of SFSS

Based on a first examination of the youth SFSS, it was indicated to collapse the five answer categories into three[5]. The recoding was done in way so that *Hardly Ever* and *Sometimes* were coded as 2 and *Often* and *Very Often* as 3 while *Never* was still coded as 1. The recoding was done using the corresponding function in Winsteps®. The following

[5] Technically the Rasch model is a type of logistic regression and, thus, susceptible to the problems of nested designs (in this case the randomized block design). However, the development of multi-level Rasch models is still very new and their use is not common practice (see Pastor, 2003 for a discussion of this topic). The experience with these types of models is very limited, so that it is difficult to judge what the appropriate approach would be. The most popular Rasch modeling software, Winsteps®, does not have a function for multi-level modeling. Taking this into account as well as that the fact that design effect is not very big as I will discuss later I felt justified in my choice to use the regular rating scale Rasch model for the purpose of creating the measures for the SFSS.

results are based on the recoded data. The item characteristic curve shown in Figure 2 below demonstrates that the recoded SFSS is a well-functioning three-item scale. The goal here is that each category (1, 2, and 3) has the highest probability of being endorsed at some point along the person score continuum (Bond and Fox, 2001). This is accomplished for the first category at the left end of the X-Axis, for the second in the middle, and for the third on the right end. The item thresholds are at -.93 and .93 logits, which is where the curves cross. That means, for example, a person with an ability score of -.93 logits is just as likely to endorse category 1 (*Never*) as they would endorse category 2 (*Hardly Ever/Sometimes*).

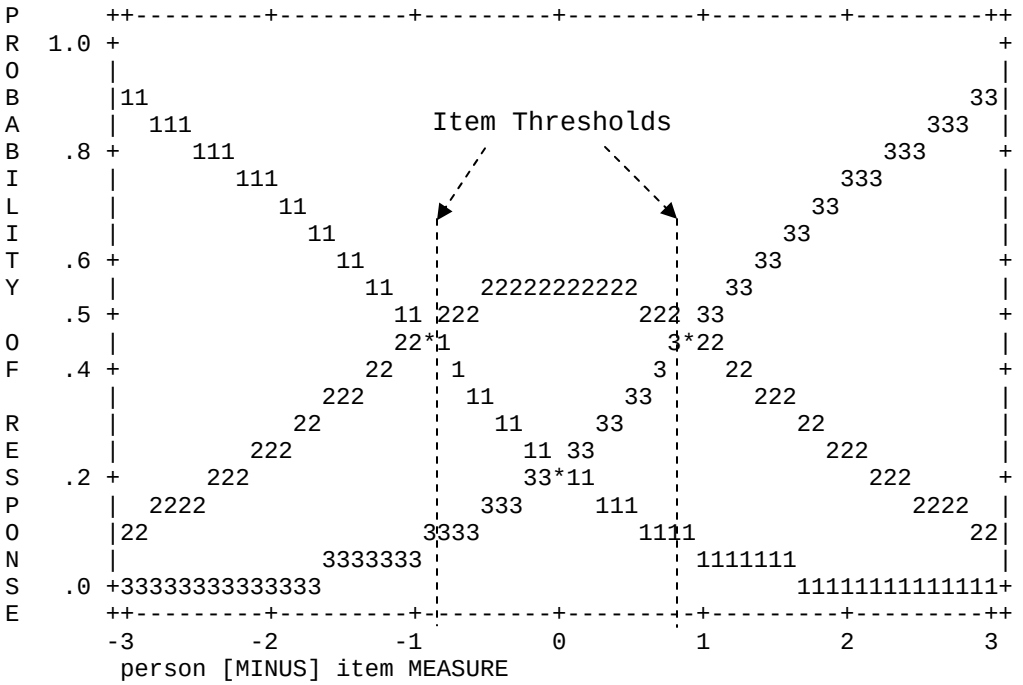

Figure 2: Youth Item Characteristic Curve

The interpretation of the results usually begins by examining the information presented in Tables 4 and 5, which provide a summary of the key model fit statistics[6]. The first statistics to consider are the person and item reliability scores. The person reliability of .91 suggests that the scale discriminates well between persons. The person reliability is approximately equivalent to coefficient alpha so that values above .80 are considered satisfactory (Clark & Watson, 1995). The item reliability of .99 indicates that the items in the youth version of the SFSS create a well-defined variable.

Table 4: Summary of 424 Measured Youths

	RAW SCORE	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	58.8	-.63	.31	1.01	-.1	1.05	-.1
S.D.	12.3	1.15	.09	.40	1.9	.65	1.7
MAX.	91.0	2.81	1.01	2.28	4.6	9.90	4.7
MIN.	34.0	-4.64	.27	.19	-6.2	.21	-5.8

REAL RMSE .35 ADJ.SD 1.10 SEPARATION 3.12 person RELIABILITY .91

Table 5: Summary of 33 Measured Items (Youth)

	RAW SCORE	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	755.8	.00	.08	1.02	.0	1.05	.0
S.D.	107.1	.79	.01	.18	2.5	.33	2.3
MAX.	917.0	2.09	.12	1.48	4.7	2.61	6.4
MIN.	512.0	-1.08	.08	.79	-3.7	.77	-3.6

REAL RMSE .09 ADJ.SD .78 SEPARATION 8.71 item RELIABILITY .99

[6] Please note that seven youths were excluded because of their extreme ratings. Extreme ratings such as rating all items "never" (which those seven did) cannot be used in Joint Maximum Likelihood Estimation (see Linacre, 2004). Later I will demonstrate that these seven extreme cases should be excluded from the all other analyses as well.

The next step is to evaluate the location of the persons relative to the item. The default is to set the item mean to 0, which I did. The mean person score is -.63 implying that the items are somewhat difficult for the youths to assign high scores. The person score distribution is not centered over 0 but is shifted somewhat downwards to the lower scores. There are several youth with very low scores. This is not surprising since, according to the leadership at PSC, most youth in foster care remain in treatment until they turn 21 years old even if they are stable and functioning normally.

At the item level it is items v10 (*Use drugs non-medica*l), v14 (*Think about hurting yourself*), and v16 (*Drink alcohol*) that are particularly difficult to endorse. This is not unexpected given that these events are expected to be less frequent relative to the other events in the SFSS and the high social desirability to underreport for these items.

Checking the residual matrix of the expected relative to the observed scores is also common practice. There are two types of fit statistics that provide summary information about the degree of deviation and misfit: outfit and infit. Outfit is based on the conventional sum of squared standardized residuals while infit is an information-weighted sum. The outfit statistic is more sensitive to extreme observations while the weighting lessens the influence of those for the infit statistics. It is common to report the outfit and infit statistics as mean square values (MNSQ) as well as standardized values (ZSTD) which is a type of *t* statistic. If the data fit the model perfectly, the *t* statistic should have a mean of 0 and a standard deviation of 1. *T* values greater than +2 or less than -2 are generally interpreted as having less compatibility with the model than expected (Bond and Fox, 2001). That means there are several people who, based on their

general scoring behavior, should have endorsed this item in a certain expected way, but did not. There are no hard-and-fast rules for interpreting the MNSQ. However, Wright and Linacre (1994) suggest as a reasonable item mean square range for rating scales is .6 to 1.4 (the potential range is 0 to infinity) with the expected value being 1.

Using these general guidelines it is clear that both the general item and person fit are very good for the youth SFSS (see Tables 4 and 5). The person infit MNSQ is 1.01 (SD=.4) and the ZSTD is -.1 (1.9) while the respective outfit values are 1.05 (.65) and -.1 (1.7). The outfit MNSQ of 1.05 indicates that there is only 5% more noise than modeled. The standard deviations are a little bit higher than the expected value which is due to some unusual observations. The item fit indices are as follows: infit MNSQ = 1.02 (.18) ZSTD= 0 (2.5) and the corresponding outfit values are 1.05 (.33) and 0 (2.3). Again, the standard deviations are a little higher than in the ideal situation.

Another aspect of the fit of the data to the Rasch model is unidimensionality. While it is clear that empirical data are always manifestations of more than one latent dimension, in the Rasch measurement framework it is important to demonstrate that there is only one primary latent variable that is represented by the measurement model[7]. Here I am applying Linacre's (1998) approach to check for multidimensionality. Using simulation studies Linacre showed that constructing Rasch measures from observational data and then conducting a principal component factor analysis of the standardized residuals is an effective way of detecting multidimensionality (see also Smith, 2004 for further investigations of the reliability of this approach). This approach is implemented in

[7] Some attempts to develop a multidimensional Rasch measurement model have been made (e.g., Briggs and Wilson, 2004), however, to what degree that fits with the Rasch measurement philosophy is still being debated.

the Winsteps® software and, thus, easily accessible. A first preliminary step is to evaluate the corrected item-total correlations for each item. As can be seen in Table 6, no item has a negative correlation and only two items have an item-total correlation below .4. Again, these items are the extreme items v16 and v10. The remaining correlations are between .40 and .63 which is a first indicator of unidimensionality according to the Rasch measurement model. A next preliminary step is to investigate the item level outfit statistics for unusual patterns that may point to multiple dimensions. Using the MNSQ there are two items (v10 and v16) that are outside the suggested range of .6 to 1.4 and 11 additional items that are significantly below or above -2/+2 for the ZSTD. It is not unusual to find items with poor fit in a longer scale of a complex latent construct. If items several items show some deviation of the empirical data from the Rasch model, the next step is to check whether there are some important connections among those items that would suggest another important factor beside the main factor represented by the measurement scores. In studying the extreme items there is no obvious connection among them that would point to another substantive dimension. As mentioned previously, except for the two extreme ones, all items have positive item-total correlations greater than .4, which would suggest that they are connected to the primary dimension in a satisfactory way.

This is further confirmed by the principal component analysis (without rotation) of the Rasch measurement residuals. The results of this analysis support the assumption that the SFSS represents one primary dimension. The measure explains 59.4% of the variance while the next contrast accounts for only 4.3% of the total variance and 10.5% of the variance unexplained by the measure. The Scree plot in Figure 3 further illustrates

these results. In investigating the loadings of the items on the second and third contrast there is not trend or systematic relationship discernable that would suggest another meaningful dimension[8]. To summarize, for the purpose of the current study it is reasonable to assume that the SFSS represents a one-dimensional scale.

All of the above investigations provide evidence that the current data fit a Rasch measurement model well and that the assumptions for using Rasch measurement person scores for the youths have been met. In addition, it was shown that the youth SFSS has good reliability and represents a clear one-dimensional scale. Next, I will investigate whether the data from the caregiver version of the SFSS has similarly good fit with the Rasch model.

[8] The actual loadings are not included in this dissertation but can be requested from the author.

Table 6: Item Summary Statistics (Youth)

ITEM	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTBIS CORR.
16	512	423	2.09	.12	1.40	3.7	2.61	6.4	.18
10	523	421	1.92	.11	1.48	4.7	1.61	3.2	.25
2	825	421	-.49	.08	1.22	3.6	1.34	4.7	.44
33	649	420	.69	.09	1.34	4.7	1.28	2.8	.43
20	772	422	-.15	.08	1.25	4.0	1.29	3.8	.42
32	629	417	.81	.09	1.18	2.5	1.05	.5	.52
5	767	423	-.10	.08	1.16	2.6	1.17	2.4	.41
21	755	420	-.06	.08	1.13	2.0	1.15	2.0	.40
17	682	422	.47	.08	1.11	1.7	1.10	1.1	.46
11	901	416	-1.04	.08	1.07	1.2	1.08	1.1	.52
18	661	423	.63	.09	1.07	1.0	1.03	.3	.51
29	737	420	.06	.08	1.06	.9	1.06	.9	.55
27	722	422	.19	.08	1.05	.8	.97	-.4	.60
13	637	423	.81	.09	1.04	.7	.97	-.3	.48
26	818	420	-.47	.08	1.04	.6	1.03	.5	.57
14	569	415	1.32	.10	1.02	.3	.91	-.7	.54
1	724	424	.20	.08	1.01	.1	.97	-.4	.50
22	767	422	-.11	.08	.97	-.4	.98	-.2	.54
15	669	423	.57	.09	.98	-.2	.89	-1.2	.60
8	761	423	-.06	.08	.89	-1.9	.98	-.3	.51
30	769	423	-.12	.08	.96	-.7	.97	-.5	.52
19	728	419	.10	.08	.97	-.5	.97	-.4	.54
9	889	422	-.89	.08	.90	-1.7	.92	-1.2	.55
28	733	419	.08	.08	.92	-1.4	.88	-1.6	.62
31	867	420	-.78	.08	.88	-2.1	.88	-1.9	.55
4	913	419	-1.08	.08	.84	-2.8	.87	-2.0	.53
23	857	421	-.70	.08	.86	-2.4	.86	-2.3	.61
7	837	423	-.55	.08	.81	-3.5	.86	-2.3	.51
25	835	422	-.55	.08	.84	-2.9	.85	-2.4	.56
3	917	422	-1.07	.08	.79	-3.7	.82	-2.8	.55
6	894	423	-.91	.08	.81	-3.5	.82	-3.0	.52
24	773	420	-.18	.08	.80	-3.6	.77	-3.6	.63
12	851	423	-.64	.08	.80	-3.7	.78	-3.5	.63
MEAN	755.8	421.1	.00	.08	1.02	.0	1.05	.0	
S.D.	107.1	2.1	.79	.01	.18	2.5	.33	2.3	

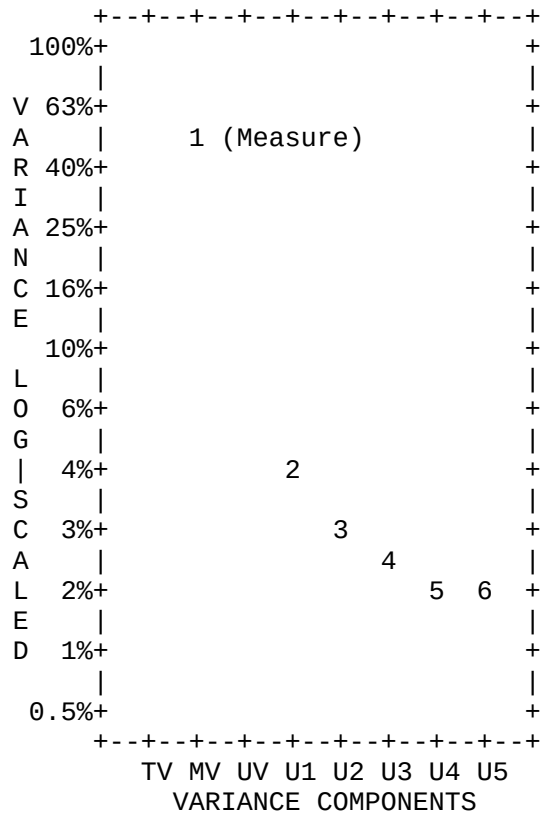

Figure 3: Variance Component Scree Plot (Youths)

Caregiver Version of SFSS

Similar to the youth version, it was indicated to recode the caregiver version of the SFSS based on a preliminary analysis. I used the same recoding scheme as I did for the youth version. The item characteristic curve in Figure 4 suggests that the recoded caregiver SFSS is a well-functioning three-item scale. The item thresholds are -1.41 and 1.41 indicating that the middle category is more likely to be endorsed by the caregivers than it is by the youths.

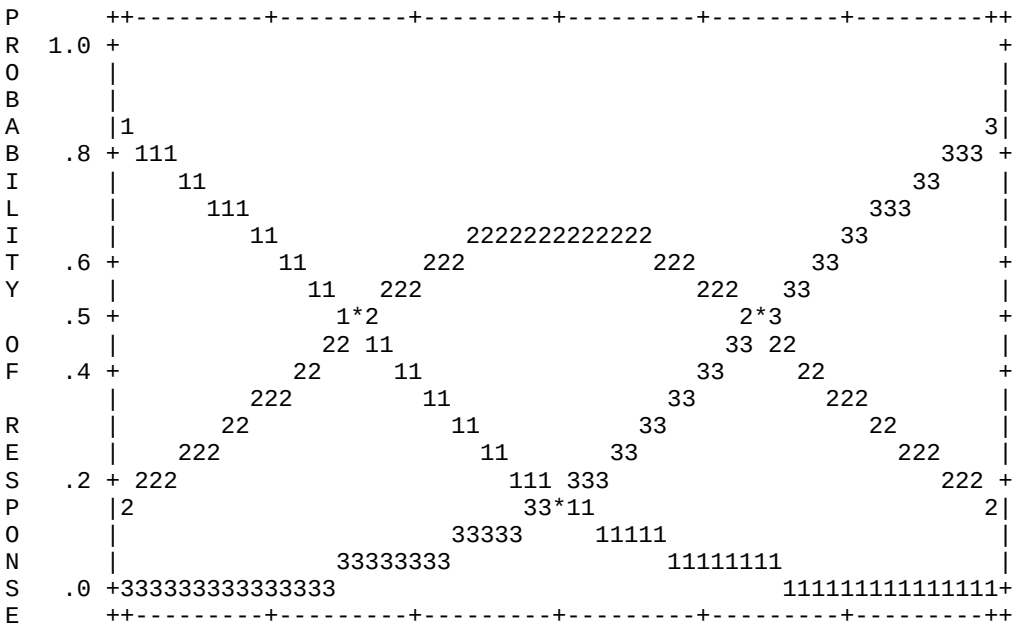

Figure 4: Caregiver Item Characteristic Curve

The person and item reliabilities are high with .92 and .99 (see Tables 7 and 8). The mean person measure score of -.13 and the person map of items illustrates that persons and items are pretty well lined up. There are three items that stand out at the high end of item difficulty (see Table 9): items v10 and v16 again as well as item v14 (*Think about hurting him-/herself*). The general infit and outfit values for persons and items indicate an overall good fit. The person infit MNSQ is 1 (SD=.42) and ZSTD is -.2 (1.8) while the respective outfit values are 1.02 (.48) and -.1 (1.7). Again, the deviations of the standard deviations from the ideal suggest that there are some caregivers and items that do not fit perfectly with the model which is expected. The goal is to evaluate whether all indicators together would suggest that the SFSS is a reasonable one-dimensional scale, which is the case.

Table 7: Summary of 323 Measured Caregivers

	RAW SCORE	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	64.8	-.13	.34	1.00	-.2	1.02	-.1
S.D.	12.0	1.33	.06	.42	1.8	.48	1.7
MAX.	93.0	3.30	1.02	2.86	5.5	3.65	5.7
MIN.	34.0	-5.24	.31	.25	-4.7	.28	-4.3

REAL RMSE .37 ADJ.SD 1.28 SEPARATION 3.50 person RELIABILITY .92

Table 8: Summary of 33 Measured Items (Caregiver)

	RAW SCORE	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	634.1	.00	.11	1.02	-.1	1.02	-.4
S.D.	92.9	1.06	.01	.25	2.9	.28	2.6
MAX.	753.0	3.20	.15	1.69	5.5	1.88	4.9
MIN.	377.0	-1.25	.10	.63	-6.0	.61	-6.0

REAL RMSE .11 ADJ.SD 1.06 SEPARATION 9.30 item RELIABILITY .99

In regard to checking for multi-dimensionality there are five items that have an item-total correlation of less than .4 although three of those are close to that value (.37, .38, and .39). The latter items are v22 (*feel nervous/shy around people*), v20 (*sleep a lot more than usual*), and v21 (*hang out with kids who get into trouble*) respectively. The items with the lower item-total correlations are the two extreme items v10 (.16) and v16 (.13). There seems no obvious connection between these items that would suggest another substantial dimension besides the main one. This is confirmed by the principal component analysis of the residuals. A large portion (70.5%) of the variance in the sample is explained by the main measure scores. The next contrast accounts for only 3.9% of the total variance and 13.1% of the residual variance. While there are multiple

items that are outside of the recommended range for item infit and outfit, the results above suggest that this does not point to another secondary dimension that would question the one-dimensionality of the caregiver SFSS. In further revisions of the instrument the developers of the SFSS may want to either recalibrate some of these items with unsatisfactory fit or consider dropping some of them. However, it is important not to rely purely on the empirical data but to also take the theoretical relevance of each item into account. These items may be considered important indicators of a youth's mental health status and should not be deleted even though they do not align perfectly with the other items in the scale. In general, it can be concluded that based on the evaluation presented above, the caregiver version of the SFSS meets the requirement for Rasch modeling and can be considered a reliable one-dimensional scale with interval level properties after it has been transformed to measure scores.

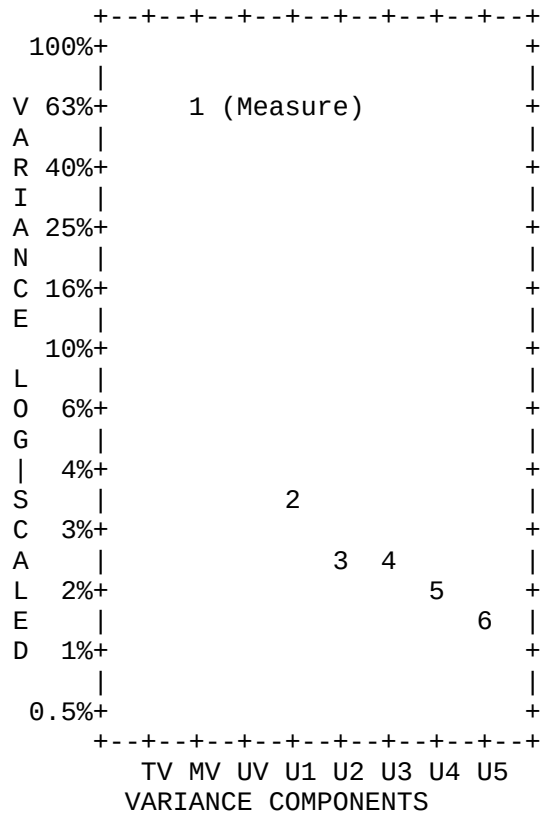

Figure 5: Variance Component Scree Plot (Caregivers)

Table 9: Item Summary Statistics (Caregiver)

ITEM NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTBIS CORR.
16	390	322	3.09	.15	1.65	5.4	1.88	3.0	.13
10	377	316	3.20	.15	1.69	5.4	1.69	2.3	.16
21	579	319	.49	.10	1.45	5.5	1.44	4.6	.39
2	648	317	-.24	.10	1.33	4.1	1.42	4.9	.41
22	604	323	.33	.10	1.18	2.4	1.32	3.5	.37
20	605	323	.32	.10	1.20	2.6	1.25	2.8	.38
5	626	321	.07	.10	1.20	2.6	1.18	2.2	.41
33	534	313	.87	.11	1.17	2.2	1.11	1.1	.48
29	594	322	.42	.10	1.15	2.0	1.14	1.6	.49
1	590	321	.44	.10	1.11	1.4	1.14	1.6	.52
27	580	322	.55	.10	1.08	1.1	1.04	.5	.52
13	609	318	.17	.10	1.06	.9	1.02	.3	.62
14	475	319	1.69	.11	1.05	.7	1.03	.3	.46
18	620	322	.14	.10	1.05	.7	1.01	.2	.56
26	684	317	-.63	.10	1.03	.4	1.01	.2	.58
15	581	321	.53	.10	.98	-.3	.95	-.5	.57
8	714	320	-.89	.10	.95	-.7	.97	-.4	.56
32	586	322	.50	.10	.96	-.5	.93	-.9	.57
30	708	321	-.79	.10	.94	-.9	.90	-1.3	.65
31	730	322	-1.01	.10	.93	-1.0	.88	-1.6	.69
25	682	320	-.54	.10	.91	-1.3	.90	-1.4	.56
24	690	322	-.59	.10	.88	-1.7	.85	-2.1	.69
11	713	321	-.85	.10	.86	-2.1	.88	-1.6	.52
4	712	320	-.87	.10	.87	-1.9	.86	-1.8	.58
17	583	320	.47	.10	.86	-2.1	.84	-2.0	.56
23	735	323	-1.03	.10	.84	-2.3	.82	-2.5	.65
19	651	318	-.28	.10	.82	-2.6	.83	-2.3	.68
7	753	322	-1.25	.11	.83	-2.6	.81	-2.5	.66
9	737	319	-1.15	.11	.80	-3.0	.75	-3.4	.65
3	723	322	-.93	.10	.71	-4.4	.76	-3.4	.57
6	744	321	-1.18	.11	.75	-3.9	.71	-4.1	.66
28	650	322	-.17	.10	.69	-4.8	.71	-4.1	.60
12	718	322	-.88	.10	.63	-6.0	.61	-6.0	.74
MEAN	634.1	320.4	.00	.11	1.02	-.1	1.02	-.4	
S.D.	92.9	2.2	1.06	.01	.25	2.9	.28	2.6	

CHAPTER III

ANALYSES

Based on the review of the relevant literature and the arguments presented in the introductory chapter I expect to find no meaningful differences between the reference group (2-week reference period) and the comparison group (3- or 6-month reference period). I hypothesized that there would be no meaningful difference in regard to four factors: 1) central tendency (as measured by the mean), 2) distribution (as measured by the proportion of youths in the clinical or high-severity range), 3) concurrent validity (as measured by the correlation with the CBCL/YSR), and 4) reliability (as measured by Cronbach's alpha). Thus, in all four cases I need to be able to demonstrate with a predefined level of certainty (e.g., 95%) that the observed differences between the groups are not greater than a pre-defined threshold below which one would consider the differences to have little to no practical relevance.

Fortunately, Rogers, Howard, and Vessey (1993) and later Searman and Seal (1998) as well as Stegner, Bostrom, and Greenfield (1996) introduced equivalence testing to social scientists including psychologists. Equivalence testing has been a popular method in the biomedical field for establishing bioequivalence (Anderson & Hauck, 1983, Berger & Hsu, 1996, Schuirmann, 1987, Westlake, 1976). Two drug formulations or treatments are considered bioequivalent if their effects on several blood concentration variables are equivalent according to a pre-defined measure of practical equivalence. It is often used in the pharmaceutical industry to demonstrate that a generic drug is

bioequivalent to a brand-name product. Regulatory agencies such as the U.S. Food and Drug Administration (FDA) allow companies to market a new generic product without clinical trials if bioequivalence has been established.

On first sight equivalence testing seems to be contrary to what every student of statistics 101 learns about hypothesis testing: One cannot prove the null hypothesis of no difference. In hypothesis testing one tries to reject the null hypothesis in favor of the alternative hypothesis stating that there are significant differences. If one fails to reject the null hypothesis, it would be inappropriate to conclude that there are no differences (Hays, 1994). In equivalence testing, however, the goal is not to demonstrate that the difference between two population parameters is zero but rather to show that the difference is smaller than what would be considered a meaningful difference.

Several approaches to equivalence testing have been developed over the years (Anderson & Hauck, 1983; Berger & Hsu, 1996; Schuirmann, 1987; Westlake, 1976). Among these, the two one-sided test procedure by Schuirmann (1987) is the most popular and well-established one (Rogers et al., 1993). Another popular approach is Westlake's Confidence Interval Procedure (Seaman & Searl, 1998; Westlake, 1976, 1981). I will briefly introduce both of these approaches.

Schuirmann's Two One-Sided Test

Schuirmann's equivalence test procedure is done in two steps: first one defines equivalence and then conducts two simultaneous one-sided hypothesis tests. Defining the equivalence depends on the researcher's understanding of what would be considered a meaningful difference between two population parameters θ_1 and θ_2 (e.g., the difference between two population means μ_1 and μ_2). That is, the investigators state *a priori* that

they will consider θ_1 and θ_2 equivalent if they differ by less than some δ in both a negative (δ_1) and positive direction (δ_2) (Rogers et al., 1993; Schuirmann, 1987). The standard for δ established by the FDA as well as the European Community, for accepting bioequivalence is a difference of about 20%. More specifically the guidelines propose that bioequivalence has been demonstrated if the ratio of the two group means is in the range of 80-125% (Wellek, 2003). However, Stegner et al., (1996) question whether this 20% level will achieve consensus within the psychological community. Wellek (2003) also questions whether the 20% standard may be too liberal in most cases. Later, I will discuss the appropriate equivalence threshold for each hypothesis separately.

Once the equivalence threshold has been defined, two one-sided hypothesis tests are conducted. The first test seeks to reject the null hypothesis ($H_{0\text{-}1}$) stating that the difference between θ_1 and θ_2 (= Δ)[9] is less than or equal to the smaller delta δ_1. The alternative hypothesis ($H_{a\text{-}1}$) is that Δ is larger than δ_1. Accordingly, the second test seeks to reject the null hypothesis ($H_{0\text{-}2}$) that Δ is greater than or equal to the larger delta δ_2. Here, the alternative hypothesis ($H_{a\text{-}1}$) is that Δ is smaller than δ_2. Where $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are the sample estimates of the population parameters (e.g., the sample means). That is:

$$\text{Test 1} \begin{cases} H_{0-1} : \Theta_1 - \Theta_2 \leq \delta_1 \\ H_{a-1} : \Theta_1 - \Theta_2 > \delta_1 \end{cases} \tag{2}$$

$$\text{Test 2} \begin{cases} H_{0-2} : \Theta_1 - \Theta_2 \geq \delta_2 \\ H_{a-2} : \Theta_1 - \Theta_2 < \delta_2 \end{cases} \tag{3}$$

[9] Note that in many cases the interval is symmetrical, that is, δ_1 is set equal to $-\delta_2$.

The corresponding test statistics are:

$$z_1 = \frac{(\hat{\Theta}_1 - \hat{\Theta}_2) + \delta_1}{S_{\hat{\Theta}_1 - \hat{\Theta}_2}} \tag{4}$$

$$z_2 = \frac{(\hat{\Theta}_1 - \hat{\Theta}_2) - \delta_2}{S_{\hat{\Theta}_1 - \hat{\Theta}_2}} \tag{5}$$

Where $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are the sample estimates of the population parameters (e.g., the sample means) and $S_{\hat{\Theta}_1 - \hat{\Theta}_2}$ is the pooled standard error of those estimates. The goal is to demonstrate statistically that the difference $\hat{\Theta}_1$ - $\hat{\Theta}_2$ is too large to have come from a distribution with a population parameter of δ_1 and simultaneously to small to have come from a distribution with a population parameter δ_2. If this has been demonstrated, it is concluded that the distribution came from somewhere in the middle, with the true difference Δ less than the meaningful difference that was set a priori by the researcher.

It is important to note that even though two tests are conducted simultaneously the Type I error rate is still equal to the alpha level (α) set by the investigator for a one sided test (e.g., α=.05). This is due to the fact that even though both $H_{0\text{-}1}$ and $H_{0\text{-}2}$ must be rejected, only one test needs to be performed (Rogers et al., 1993). The test that has the shorter distance between $\hat{\Theta}_1$ - $\hat{\Theta}_2$ and either δ_1 or δ_2 will result in the smaller test statistic and consequently the larger *p* value of the two possible tests. If the null hypothesis of the test with the larger *p* value is rejected it follows that the null hypothesis of the other test will be rejected as well because of its smaller p value. If, on the other hand, the first test fails to reject the null hypothesis, the second test does not need to be performed because

both null hypotheses need to be rejected to infer that Δ is within the bounds of δ_1 and δ_2. Consequently, only one statistical test with a type I error rate α needs to be performed. Rogers et al., (1993) illustrate that Schuirmann's method can be used with a variety of population parameters such as means, proportions, and effect sizes.

Westlake's Confidence Interval Procedure

Westlake's (1976, 1981) confidence interval approach is a helpful supplement to Schuirmann's method because of the way the test can be graphically displayed. Similar to the approach described above, a lower (δ_1) and upper threshold (δ_2) for a meaningful difference are defined prior to conducting the test. It can be demonstrated that one can be 100(1-α)% confident that Δ is within the bounds of δ_1 and δ_2 , if the 100(1-2α)% confidence interval

$$[(\hat{\Theta}_1 - \hat{\Theta}_2) - z_{1-\alpha/2} S_{\hat{\Theta}_1 - \hat{\Theta}_2} \ , (\hat{\Theta}_1 - \hat{\Theta}_2) + z_{1-\alpha/2} S_{\hat{\Theta}_1 - \hat{\Theta}_2}]$$

is within the those equivalence bounds (Rogers, et al., 1993; Seaman & Searl, 1998). That is, the equivalence test is at the α-level, even though 2α is used to construct the interval. The logic here is parallel to the argument presented above for the confidence level in the case of the two one-sided tests (Rogers, et al., 1993). Assuming an α-level of .05, Figure 6 illustrates a case where the confidence interval is not completely within the bounds of δ_1 and δ_2 and one would, thus, not be able to conclude with 95% confidence that the parameters (e.g., the group means) are practically equivalent. Figure 7, however, shows a case where the confidence interval is completely within the bounds of the equivalence criteria and, thus, practical equivalence can be inferred.

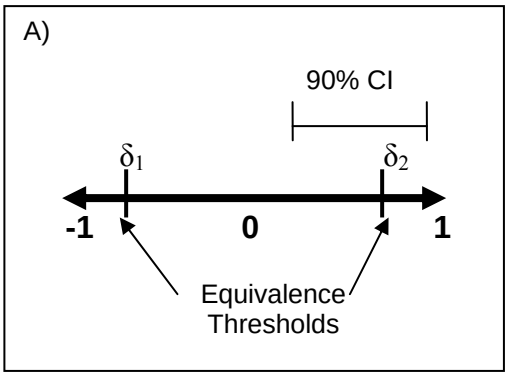

Figure 6: 90% Confidence Interval (CI) not Completely Within the Equality Bounds

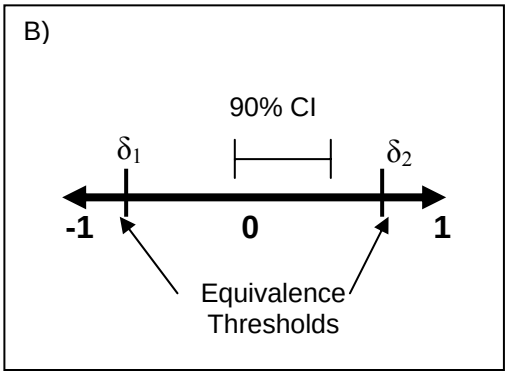

Figure 7: 90% CI Completely Within the Equality Bounds

Power Analyses for Equivalence Tests

The power of a statistical test is the probability that the null hypothesis, H_0, will be rejected when the alternative hypothesis, H_1, is true. Since in equivalence testing the alternative hypothesis is that there are no meaningful differences, the statistical power of the test is the probability of inferring equivalence, if indeed the population parameters (e.g., the population means) are not different in a meaningful way as determined by the lower and upper limit δ_1 and δ_2. That is, given that the difference of the population parameters is within the equivalence interval, the power p is the probability P that the test statistic z_1 and $-z_2$ are equal or greater than the critical value of $z_{(1-\alpha, df)}$ (Phillips,1990).

Prior to conducting the analyses I used SAS to assess statistical power for hypotheses 1a and 1b as well as 2a and 2b. In all four cases there was sufficient power ($\beta \geq .8$) to establish statistical equivalence within reasonable equivalence bounds. The only exception is when the observed difference is very close to or beyond the equivalence threshold. In that case the equivalency test is not a powerful test and one should use traditional significance testing instead. For calculating the power for the test of hypotheses 3a and 3b I used a procedure developed to assess the power for equivalence test of two correlations recently presented by Miriam Kraatz (2006). Because of the smaller sample size available for testing this hypothesis (n=129 for the youths and n=111 for the caregivers), the results of the power analysis were negative, that is, there is not sufficient power to establish equivalence within any reasonable bounds. If, for example the correlation in the reference sample would be .8 the equivalence bounds would have to be set to plus and minus .21 to have power of .8 to detect equivalence. That is, one would have to assume that a difference of .21 in the correlation is not of practical significance, which is unreasonable. Finally, to my knowledge, there is no existing procedure to conduct power analyses for the types of tests proposed for hypotheses 4a and b. Since the coefficient alpha is a type of correlation, I used the procedure developed by Kraatz (2006) for estimating power for equivalence tests of correlation, instead. The results of this test indicated that there is sufficient power. For other reasons, I will limit the tests for hypotheses 4a and 4b to the confidence interval approach for which pre-established power estimates are not as critical because the span of the interval itself provides an indication of the confidence one can have in the results of the test. That is, the narrower the confidence interval, the more powerful and conclusive are the inferences.

CHAPTER IV

RESULTS

Hypothesis 1—Central tendency results

For the convenience of the reader I will repeat hypothesis 1a:

> 1a) The **difference of the mean** SFSS score (as rated by the youth) between the reference and the comparison group is not meaningfully different.

Hypothesis 1b is equivalent except that it is applied to the caregiver version of the SFSS. As explained above, the first step in testing these two hypotheses is defining the equivalence threshold for a meaningful difference in the group means. What is considered a meaningful difference in a clinical or practical sense, especially in regard to the difference in the scores of mental health outcome measures from one time point to the next (that is, clinically meaningful change), has been the subject of much debate (Jacobson, et al., 1999; Thompson, 2002). As mentioned earlier a widely accepted understanding of clinically significant change in the context of psychotherapy outcome studies was proposed by Jacobson and colleagues (Jacobson, et al., 1999; Jacobson & Truax, 1991). The definition of clinically significant change is as follows: "(a) The magnitude has to be statistically reliable and (b) by the end of therapy, clients have to end up in a range that renders them indistinguishable from well-functioning people." (Jacobson, et al., 1999, p. 300) While I am not able to study change with the cross-sectional dataset available to me, I am able to apply these criteria to the differences in the reference and comparison group. I will apply the first criteria (statistically reliable

magnitude) for testing hypotheses 1a and 1b while the second criteria is related to what is being tested in hypotheses 2a and 2b as explained earlier.

The reliable change index (RC) was developed by Jacobson and colleagues (see Jacobson and Truax, 1991) to determine whether the magnitude of change for a given client is statistically reliable. It takes the reliability of the measure being used into account to provide some certainty that the observed change from one measurement instance to the next is not due to chance or the unreliability of the measure but actually reflects a true change in the status of the client. Thus, if the difference between the reference group mean and the comparison group mean on the SFSS is smaller than what would be considered a reliable change on that same measure according to Jacobson and colleagues I conclude that the two groups are equivalent. The reliable change threshold, that is, the minimum difference or change (δ) is computed as: $\delta=1.96 * S_{diff}$ for a 95% certainty that the change is not simply due to chance. The standard error of difference S_{diff} can be computed directly from the standard error of measurement:

$$S_{diff} = \sqrt{2(S_E)^2} \text{ with } S_E = S\sqrt{1 - r_{xx}} \qquad (6) \text{ and } (7)$$

S is the standard deviation of the experimental group at baseline or in this case the reference group. For the reliability r_{xx} I will use the estimates of internal reliability presented as part of the Rasch analysis earlier in this chapter.[10]

Thus, if μ_1 and μ_2 are the population means of the reference and the comparison group the hypothesis tests for hypothesis 1a and 1b are the following:

[10] If change over time is studied and an estimate of the test-retest reliability is available, that estimate is most often used in calculating the RCI. Since the value for the test-retest reliability is often smaller than the value for the internal reliability estimate, using the latter results in a more conservative equivalence test.

$$\text{Test 1} \begin{cases} H_{0-1} : \mu_1 - \mu_2 \leq -\delta \\ H_{a-1} : \mu_1 - \mu_2 > -\delta \end{cases} \tag{8}$$

$$\text{Test 2} \begin{cases} H_{0-2} : \mu_1 - \mu_2 \geq \delta \\ H_{a-2} : \mu_1 - \mu_2 < \delta \end{cases} \tag{9}$$

And the two test statistics are:

$$t_1 = \frac{(\bar{x}_1 - \bar{x}_2) + \delta}{S_{\bar{x}_1 - \bar{x}_2}} \quad \text{and} \quad t_2 = \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{S_{\bar{x}_1 - \bar{x}_2}} \qquad (10) \text{ and } (11)$$

$\bar{x}_1$ and $\bar{x}_2$ are the sample means of the reference and the comparison group and, if an independent random sample can be assumed, $S_{\bar{x}_1 - \bar{x}_2}$ is the pooled standard error of the means:

$$S_{\bar{x}_1 - \bar{x}_2} = \left[\left(\frac{(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2}{n_1 + n_2 - 2} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right]^{1/2} \tag{12}$$

where n_1 and s_1 are the sample size and standard deviation of the reference group while n_2 and s_2 are those of the comparison group.

Since the study used a randomized complete blocks design with regional offices as the blocks it cannot be assumed that the observations are necessarily independent. The fact that students were randomized within a regional office can influence the estimate of the standard error for the effect of treatment, that is, the variability of the estimate of the difference in reference periods as well as the estimate of the mean difference itself (Milliken & Johnson, 2002). Thus, to obtain a more precise and accurate estimate of the standard error as well as an estimate of the adjusted mean difference for the equivalence test of the two experimental conditions I fitted a mixed model to the data using SAS PROC MIXED (Littell, Milliken, Stroup, & Wolfinger, 1996). In this model the

experimental treatment variable (REPERIOD) was entered as a fixed effect while the block variable (OFFICE) was treated as a random effect. Entering the blocking variable office as a random factor controls for the intraclass correlation (ICC). The ICC for office in the youth data is .088 while it is .071 for the caregiver sample.[11]

Sometimes when pooling experimental data from multiple sites, a treatment by site interaction can be present which would require a different model and interpretation of the main effects (Worthington, 2004). This is unlikely in this case given that it was a cross-sectional study and the treatment in this case (different reference periods in a questionnaire) is unlike the types of treatments researchers are testing in typical multi-center studies where this type of interaction can occur (e.g., staff implementing a certain treatment protocol differently). This assumption of no significant interaction was supported by an empirical investigation for both the youth and the caregiver data. Thus, the interaction term was not included in the model and the model for randomized block designs suggested by Milliken and Johnson (2002) and Littell et al., (1996) was used to obtain estimates for the standard error and the adjusted mean difference. As explained above, this is a mixed model with treatment as a fixed effect and office entered as a random effect. The SAS code for estimating this model for the youth data is provided in Appendix A. The model estimates were generated using restricted maximum likelihood estimation (which is the default for SAS PROC MIXED). The estimates of the adjusted mean difference and the corresponding standard error where then entered into equations

[11]In principal the youth and caregiver scores should also be nested within counselors. In most cases in this sample, however, there were very few youth nested within counselor. The mode of the number of youths nested within counselors is one youth per counselor. In simulation studies Muthen and Satorra (1995) have shown that the influence of the nested design on the standard error and bias in estimates is negligible if the cluster sizes (< 7) and the ICC (<.1) are relatively small. In addition, only the youth and caregiver ratings are considered in this inquiry, not the counselor ratings. Thus, for the purpose of this study the counselor level is not further considered in the analysis.

10 and 11 to assess whether the mean difference is small enough to consider it not meaningful using the reliable change index threshold as the benchmark. With $\alpha = .05$ the critical test statistic is 1.96. If $|t_1|$ and $|t_2|$ are both greater than 1.96, it can be concluded that the comparison group is not different from the reference group in a meaningful way. In addition, the 90% confidence interval is calculated so that it can be compared to the equivalence bounds following the Westlake approach of establishing equivalence. I am also presenting the results of a traditional significance test as a comparison.

Outlier Analysis

In the univariate analysis of the SFSS youth measurement scores 11 observations were identified as extreme. These observations had extreme low scores[12] (-4.64 or lower), seven of which where at the bottom of the range (-5.84), that is, all items of the SFSS were endorsed as "never". The other extreme scores had at least one or two items endorsed with something different from "never." These extreme scores resulted in a high kurtosis (2.95) even though the remaining observations appeared to be approximately normally distributed. I used the diagnostic statistics for mixed models available with SAS PROC MIXED to determine to what degree these univariate extreme values would lead to outliers in the residuals and what their influence on the model estimates would be. In studying the restricted likelihood distances (a measure of overall influence), the DFFITS statistics (a measure of the effect on fitted values), and the effect of the observations on the parameter estimates of the fixed and random effects (Cook's D) it became evident that the seven most extreme values are influential especially in regard to the estimates for

[12] This is before standardization. However, the mean is close to 0 and the standard deviation close to 1.

office. A second residual analysis studying the influence of the set of values clustered in each office shows indeed that those offices which contain at least one of these seven extreme observations stand out. In addition, leaving these observations in the model leads to a spurious treatment by office interaction effect that is only due to these extreme cases and completely disappears once those observations are removed. Each of these extreme values has an externally studentized residual value of 3.8 or higher which, according to Cohen, Cohen, West, and Aiken (2003), identifies them as outliers given that the sample is of medium size. It is very likely that the youths did not answer the questions honestly and simply created a pattern by going down the left most column. Given the reasonable doubt that these extreme scores represent the true status of the youths and the diagnostic criteria provided above, I decided to trim the data and remove these seven observations from the data for any further analyses. Removing these data resulted in approximately normally distributed residuals.

The same procedure and rationale as used for the youth SFSS score was applied for the caregiver data. Two cases were identified as extreme outliers (external studentized residual equal to -4.7 and -4.3) with substantial influence (especially on the estimate of the treatment effect) and, thus, were excluded from all further analyses. In addition, one office only had two caregivers who completed the SFSS and both were in the same experimental condition. Thus, no within office mean difference could be estimated for this office. Once the cases with extreme scores were trimmed the measurement scores were standardized with a mean of 0 and standard deviations of 1 for easier interpretation.

Results for Youth

The results for the test of hypothesis 1a are presented in Table 10. The estimates for the adjusted means are -0.146 for the reference group and 0.049 for the comparison group suggesting that in this sample the level of severity is only slightly lower when the two-week reference period is used. The difference of adjusted mean differences is .195 and with a standard error of .094 indicating statistically significant differences using a traditional two-tailed *t*-test level with a 95% confidence level: $t(404) = 2.16$, $p = .032$. A statistically significant result does not necessarily imply a meaningful difference in a practical sense. That is the advantage of equivalence testing.

The equivalence thresholds using the reliable change index as explained above are -.79 and .79. Applying these thresholds to Schuirmann's two one-sided tests demonstrates statistically significant equivalent group means. This can be inferred from the fact that the smaller test statistic with the larger p-value for $\alpha = .05$ is: $t(404) = 6.33$, $p < .001$. Furthermore, the 90% confidence interval (.040 to .349) is clearly within the equivalence bounds (-.79 and .79) which is illustrated in Figure 8.

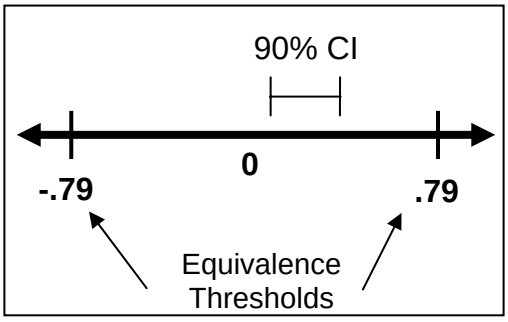

Figure 8: Results for Differences in Mean Youth SFSS Scores

Table 10: Results for Hypothesis 1

Traditional and Equivalence Test Results for SFSS Scores Mean Differences (2 Weeks vs. 3/6 Months)
(YOUTH: *df* = 404; Equivalence criteria = ±.79; CAREGIVER: *df* = 303; Equivalence criteria = ±.78)

	Adj. Means		Difference		Traditional (2-tailed)				Equivalence			
							95% CI				90% CI	
	2 W	3/6 M	M_{diff}	*SE*	*t*	*p*	LCL	UCL	*t*	*p*	LCL	UCL
YOUTH	-.146	.049	.195	.094	2.16	.032†	.019	.400	6.33	<.001*	.040	.349
CAREGIVER	-.182	.209	.391	.107	3.64	<.001†	.180	.603	3.67	<.001*	.215	.567

Note. * $p < .05$ for equivalency, per each one-tailed test. Only the p value for the smaller t value is listed.
† $p < .05$ for traditional test, two-tailed

Table 11: Results for Hypothesis 2

Traditional and Equivalence Test Results for Differences in Proportions of Youths who are in the High Score Range
(YOUTH: Equivalence criteria = ±.10; CAREGIVER: Equivalence criteria = ±.10)

	Proportions (clinical)						Traditional (2-tailed)				Equivalence			
	2 Weeks		3/6 Months		Difference				95% CI				90% CI	
	p	*n*	*p*	*n*	Dif.	*SE*	*z*	*p*	LCL	UCL	*z*	*p*	LCL	UCL
YOUTH	.23	229	.34	195	.11	.044	2.62	.009†	.029	.200	0.33	.371	.043	.186
CAREGIVER	.21	177	.33	146	.13	.057	2.23	.026†	.011	.240	0.49	.312	.033	.222

Note. * $p < .05$ for equivalency, per each one-tailed test. Only the p value for the smaller t value is listed.
† $p < .05$ for traditional test, two-tailed

Result for Caregivers

I found similar results for the caregiver data which can also be found in Table 10. The estimated difference between the adjusted mean for the reference group (-0.182) and the one of the comparison group (0.209) was with 0.391 (*SE* = .107) somewhat larger than the comparable one for the youth scores. This difference is statistically significant using the traditional two-sided *t*-test $t(303) = 3.64$, $p < .001$. However, the equivalence test is also statically significant with $t(303) = 3.67$, $p < .001$ for the smaller test statistic of the two one-sided tests. Thus, with 95% confidence, we can conclude that the difference between the caregiver means is less than the pre-established equivalence criterion of .78. The 90% confidence interval (.215 to .567) is within the equivalence bounds of ±.78 as illustrated in Figure 9.

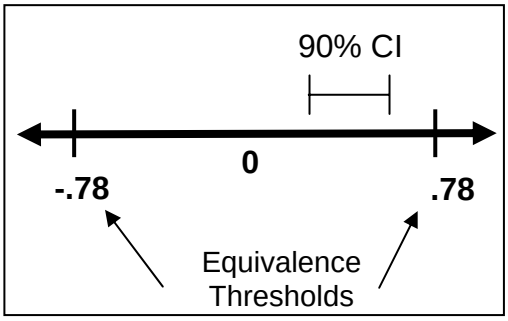

Figure 9: Results for Differences in Mean Caregiver SFSS Scores

Hypothesis 2—Variability results

Hypothesis 2a was stated as:

> 2a) The proportion of youths who are **in the high severity range** (based on the youths' ratings) in the comparison group is not meaningfully different of that in the reference group.

Again, hypothesis 2b was the same applied to the caregiver version of the SFSS. Whether a youth is in the high severity range will be determined according to the instructions in the manual of the SFSS (Bickman, et al., 2006). The manual suggests that a youth with a standardized *T*-score rating of 63 is in the high severity range, if the rating came from the youth. If the rating was done by the caregiver, the respective cut-off value is 73. According to the manual the classification of high severity on the SFSS is comparable to the "clinical range" classification of the CBCL and YSR. Since the cut-off value is a standardized *T*-score that is based on the raw scores, I will use the raw scores for testing this hypothesis rather than the measure scores used for hypothesis 1.

The test for this hypothesis concerns the equivalence of proportions. Thus, it is similar to the procedures for assessing baseline equivalence as illustrated in Example 3 in Rogers et al., (1993). These authors applied an equivalence criterion of 20%. That is, they inferred equivalence if the proportion of the comparison group was within 20% of the reference group proportion. According to Wellek (2003), however, "everyday experience shows that most people will rate probabilities of medium size differing by no more than 10%, as rather similar; 20% or more is usually considered indicating a different order of magnitude in the same context." (p. 12). If the reference proportion is .5, the two approaches would yield the same criterion because 20% of .5 is .1, that is 10%. For smaller proportions the latter approach is more lenient (that is a wider equivalence

interval) while for proportion greater than .5 the latter approach would be more conservative. For the purpose of testing hypotheses 2a and 2b I will use the latter approach.

Thus, if p_1 is the proportion of youths in the high severity range in the reference population and p_2 is the corresponding parameter in the comparison population, the minimum difference δ is equal to .1. According to Rogers et al., (2003) the two one-sided tests are as follows:

$$\text{Test 1} \begin{cases} H_{0-1} : p_1 - p_2 \leq -\delta \\ H_{a-1} : p_1 - p_2 > -\delta \end{cases} \tag{13}$$

$$\text{Test 2} \begin{cases} H_{0-2} : p_1 - p_2 \geq \delta \\ H_{a-2} : p_1 - p_2 < \delta \end{cases} \tag{14}$$

The two corresponding test statistics are:

$$z_1 = \frac{(\hat{p}_1 - \hat{p}_2) + \delta}{S_{\hat{p}_1 - \hat{p}_2}} \quad \text{and} \quad z_2 = \frac{(\hat{p}_1 - \hat{p}_2) - \delta}{S_{\hat{p}_1 - \hat{p}_2}} \qquad (15) \text{ and } (16)$$

The sample estimates $\hat{p}_1$ and $\hat{p}_2$ are the proportions of youths in the high severity range in the reference and the comparison sample. If n_1 and n_2 are the sample sizes of the reference and the comparison group, the pooled standard error is computed as follows:

$$S_{\hat{p}_1 - \hat{p}_2} = \left[\left(\frac{\hat{p}_1 (1 - \hat{p}_1)}{n_1} \right) + \left(\frac{p_2 (1 - \hat{p}_2}{n_2} \right) \right]^{1/2} \tag{17}$$

Like before, if |z1| and |z2 | are both greater or equal to 1.96 for α = .05, the null hypothesis of both test 1 and test 2 is rejected and equivalence in regard to the proportions of youth in the high severity range for the reference and the comparison group can be inferred.

Results for Youth

According to the youth self-reports forty-five out of 195 youths (23%) are in the high severity (clinical) range in the reference group while in the comparison group 79 out of 229 youths (34%) are in that range, a difference of 11% (see Table 11). This is more than the 10% threshold. As explained earlier, the power of equivalence tests drops close to 0 if the observed differences surpasses the equivalence threshold. Thus, it comes to no surprise that the observed difference of .11 is not statistically equivalent within the .1 bounds ($z = 0.33$, $p = .371$). Figure 10 illustrates that the 90% confidence interval is not fully contained by the equivalence bounds. The traditional two-tailed test for differences, however, is significant: $z = 2.62.$, $p = .009$.

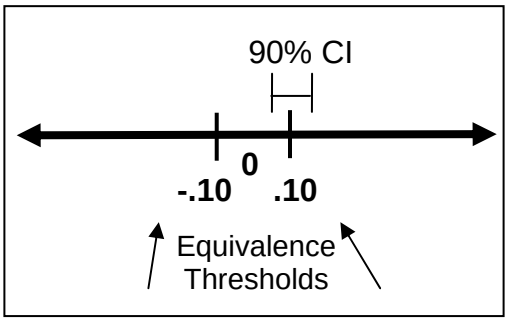

Figure 10: Results for Differences in Proportions (Youth)

Results for Caregivers

The results for the caregivers are almost identical with very similar proportions. According to the caregiver report, 30 out of 146 youths (21%) are in the high severity (clinical) range in the reference group whereas in the comparison group 59 out of 177 youths (33%) are in that range, a difference of 13%. One cannot conclude with 95% certainty that this difference is within the equivalence bounds ($z = 0.49$, $p = .312$) as illustrated in Figure 11. The results of the traditional two-tailed test suggest that the difference is statistically significant ($z = 2.23$, $p = .026$).

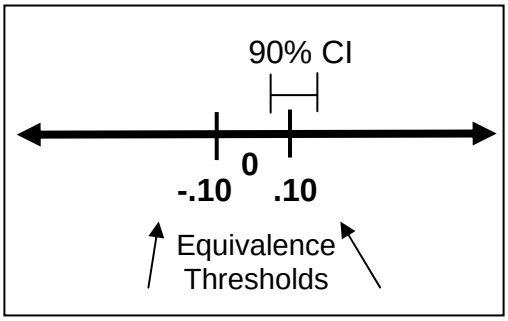

Figure 11: Results for Differences in Proportions (Caregivers)

Hypothesis 3—Validity results

Hypothesis 3a has been formulated as:

3a) The **correlation** of the SFSS scores (rated by the youth) with the YSR in the comparison group is not meaningfully different from the respective correlation in the reference group.

Hypothesis 3b is the same for the caregiver version of the SFSS and the CBCL instead of the YSR. The correlation of the youth SFSS and the YSR total score is .73 (N = 61) in the reference sample and .79 (N = 68) in the comparison group. The correlation of the caregiver ratings of the SFSS and the CBCL total score is a generally higher with .86 (N = 55) in the reference sample and .88 (N = 56) in the comparison group. While the differences between these correlations are small, the current sub-samples that were available for testing the hypotheses are too small to be able to conclude with any reasonable certainty that the two correlations are equal within any meaningful equivalency thresholds. Consequently, no statistical tests are being conducted.

Hypothesis 4—Internal reliability results

Hypothesis 4a has been formulated as:

> 4a) The estimate of the **Internal Reliability** of the SFSS scores (rated by the youth) assessed in the comparison sample is not meaningfully different from the respective reliability estimate in the reference group.

Hypothesis 4b is the same for the caregiver version of the SFSS. Several measures of reliability have been developed such as test-retest, the split-half, and Cronbach's coefficient alpha. The latter, first introduced by Cronbach in 1951, is probably the most popular measure of internal reliability in the social sciences. It is a special type of intraclass correlation describing the reliability of a test that is the sum of *k* items (Bonett, 2003). Feldt (1969), and Alsawalmeh and Feldt (1994) have proposed approximate *F* tests for testing the equivalence of two coefficient alphas. More recently, Bonett (2003) introduced a test that uses a *z*-statistic and is based on the normal distribution. While these authors claim to test for equivalence, their proposed tests are actually traditional

significance tests with the null-hypothesis stated as no difference. However, one can adapt Bonett's approach and apply it to Westlake's confidence interval approach to test for equivalence.

According to Bonett (2003) an approximate test for significant differences in two coefficient alphas (r_{xx1} and r_{xx2}) can be obtained using the following test statistic:

$$z = \frac{\ln(\hat{\lambda})}{SE_{\lambda}}, \tag{18}$$

$$\text{where } \hat{\lambda} = (1 - r_{xx1}) / (1 - r_{xx2}) \text{ and} \tag{19}$$

$$SE_{\lambda} = \left[\frac{2k_1}{(k_1 - 1)(n_1 - 2)} + \frac{2k_2}{(k_2 - 1)(n_2 - 2)} \right]^{1/2}. \tag{20}$$

Accordingly, it can be shown that

$$\exp[\ln(\hat{\lambda}) \pm z_{\alpha/2} SE_{\lambda}] \tag{21}$$

is an approximate 100(1-α)% confidence interval for λ (Bonett, 2003). Thus, using Westlake's approach one can test whether the 90% confidence interval is within the bounds of the equivalence criteria δ_1 and δ_2.

Determining what the appropriate values for δ_1 and δ_2 would be is not as straightforward as in the previous cases. Before applying the 10% or 20% rule it is helpful to study an example because the meaning of the difference in the context of reliability coefficients is not intuitive. Assume, for example, we are using the SFSS to study the correlation between symptom severity and gender. Assume that the latter is measured perfectly ($r_{yy} = 1$) while the reliability of the SFSS for the reference sample is $r_{xx1} = .92$. If we further assume that the true correlation between the symptom severity and gender is .5 and the reliability of the SFSS in the comparison sample r_{xx2} is .83 (an

approximate 10% difference) the observed empirical correlation in the reference group would be estimated to be .48[13] while in the comparison group it would be .456, a difference of approximately 5 percent. This is a small difference that in most cases in social science research would not be considered as meaningful. Thus, applying the 10% rule in the case of coefficients alpha seems appropriate assuming that the reliability estimate will be in the acceptable range of .8 to .99. Because the methods introduced in this section have been developed using the traditional estimates for coefficient alpha, I will conduct the tests using those estimates rather than the ones provided as part of the Rasch measurement model.

Results for Youths

As can be seen in Table 12 the coefficient alpha based on the youth reference sample (r_{xx1} = .921) is very similar to the one based on the comparison sample (r_{xx2} = .932). It is not surprising that this difference of .011, leading to a $\hat{\lambda}$ estimate of 1.159, is not statistically significant using a traditional two-tailed test for differences (z = 1.04, p = .299). On the other hand, one can clearly see that the 90% confidence interval (.917 to 1.464) is within the equivalence bounds of .46 and 7.90 for λ (see also Figure 12). Thus, one can conclude with 95% confidence that the true difference between the population coefficient alphas is less than 10% and, consequently, not meaningful in a practical sense.

[13] The attenuation is calculated using the following formula $r_{xy} = \rho_{xy} (r_{xx} r_{yy})^{1/2}$ (Cohen, et al., 2003).

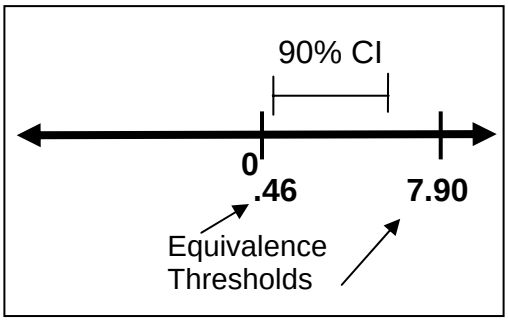

Figure 12: Results for Differences in Coefficient Alpha (Youths)

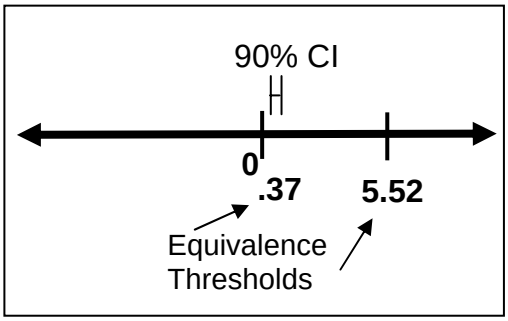

Figure 13: Results for Differences in Coefficient Alpha (Caregivers)

Results for Caregivers

The difference in coefficient alphas based on the caregiver samples is slightly larger compared to the youth samples (.17 with $r_{xx1} = .945$ and $r_{xx2} = .928$). However, the corresponding $\hat{\lambda}$ estimate of 0.761 is not statistically significant based on a traditional two-tailed test ($z = 1.66$, $p = .097$) and the 90% confidence interval (.581 to .998) is clearly within the equivalence bounds of .37 and 5.52 for λ (see also Figure 13). Consequently, one can infer that there are no meaningful differences between the two coefficient alphas.

Table 12: Results for Hypothesis 4

Traditional and Equivalence Test Results for Differences in Coefficient Alpha (r_{xx})
(Equivalence bounds for λ: YOUTH (0.46; 7.90); CAREGIVER (0.37; 5.52)

	Coefficient Alpha							Traditional (2-tailed)				Equivalence	
	2 Weeks		3/6 Months							95% CI (for λ)		90% CI (for λ)	
	r_{xx}	n	r_{xx}	n	$\hat{\lambda}$	$\ln(\hat{\lambda})$	SE	z	p	LCL	UCL	LCL	UCL
YOUTH	.921	191	.932	224	1.159	.15	.142	1.04	.299	.877	1.531	.917	1.464
CAREGIVER	.945	143	.928	168	.761	-.27	.164	1.66	.097	.552	1.051	.581	.998

CHAPTER V

DISCUSSION

In this chapter I will first interpret the results presented in the previous section. Then, I will discuss some of the inconsistency in the findings as well as the nature of the differences found. This led me to do some further explorative analyses in regard to differences by item and respondent characteristics. The most interesting findings of this explorative work are presented and discussed. Finally, I will address some of the limitations in the current study and propose some directions for future research.

Table 13 provides an overview of the findings for each hypothesis test presented above. Two things stand out. First, for hypothesis 1 both the traditional test for differences and the equivalence test are significant. That is, the means are different and equivalent according to the equivalence criteria. Second, while the hypothesis of equivalence was confirmed for means and coefficients alphas in hypotheses 1 and 4, it was not for the proportions of youth in the high severity range in hypothesis 2. As noted in Table 13, hypothesis 3 was not tested due to lack of power. Since the interpretation and discussion of these findings are related I will cover them both together.

Table 13: Summary of Results

Hypothesis	Traditional Test	Equivalence Test
1) Means		
a) Youth	Different	Equivalent
b) Caregivers	Different	Equivalent
2) Proportions High Severity		
a) Youth	Different	Not Equivalent
b) Caregivers	Different	Not Equivalent
3*) Correlations with YSR/CBCL*		
a) Youth	*NA*	*Not conducted because of lack of power*
b) Caregivers		
4) Coefficients Alpha		
a) Youth	Not Different	Equivalent
b) Caregivers	Not Different	Equivalent

Situations where the test results suggest that the population parameters are different as well as equivalent, as in the case of hypothesis 1, are not very common but they do occur. In the before mentioned article by Rogers et al. (1993), for example, the authors found 1 out of 27 (3.7%) tests conducted in their third example to be different and equivalent. The primary reason for this is that typically the traditional test seeks to reject the null hypothesis that there are no differences at all, that is, a difference of zero. Consequently, even very small differences can become significant with a large enough sample and sufficient statistical power (Hays, 1994). Equivalence testing on the other hand tries to establish that the there are no meaningful differences using an equivalence threshold that is greater than zero.

Nevertheless, a finding like this definitely warrants further investigation and discussion especially in the light of the results for hypothesis 2, where the populations were deemed different and not equivalent. The determination of equivalence depends on

the choice of the equivalence threshold. The larger the distance between the thresholds on either end, the more likely one will find the parameters (in the case of hypothesis 1 the means) to be equivalent. The thresholds selected for hypothesis 1 were based on the reliable change index developed by Jacobson and Truax (1991). This resulted in rather large equivalence thresholds of .79 for the youths and .78 for the caregiver sample. Considering that the scores were standardized prior to analysis, the standard deviations in these samples is approximately 1 leading to effect sizes of .79 and .78. According to Cohen (1992) an effect size of .8 is considered large. Thus, the requirements for establishing equivalence using the reliable change index seem rather liberal. Had I used Cohen's suggestion for a medium effect size of .5 instead, only the youth sample would render a significant result for the equivalence test. The 90% confidence interval for the estimate of the mean difference in the caregiver sample (.215 to .567) would already reach outside of the bounds of the equivalence thresholds of -.5 and .5. If we would use an even more conservative criterion, such as .2 for small effects, the equivalence tests for both, the youth and the caregiver sample, would not be significant. It is also worth mentioning that the confidence intervals in both cases (i.e., the youth and the caregiver sample) do not include zero. In both cases the longer reference period results in a higher mean score, which would suggest that, on average, both the youths and the caregivers rate the level of the youths' severity to be higher if a three- or six-month reference period is used compared to a two-week period. This difference is slightly more pronounced in the caregiver sample than in the youth sample.

That there are actually differences in the response behavior is further supported by the significant equivalence tests for hypothesis 2. The difference of 11% (youths) and

13% (caregivers) in the proportions of youth rated to be in the high severity range is statistically significant as well as of practical significance. The direction of the difference between the two experimental groups here is the same as for the means. If the longer reference period is used, more youths fall into the high severity range. Again, this difference is slightly larger for the caregiver sample. The fact that the coefficient alphas tested in hypothesis 4 are not significantly different and are also equivalent does not contradict the findings above. The first two hypotheses are investigating equivalence or differences in regard to the distribution in the two groups while a coefficient alpha is an assessment of how the items in the scale interrelate and how consistent a scale is. These are qualitatively very different things. In summarizing this discussion so far, I conclude that the SFSS measure is a consistent measure independent of the reference period used. That is, the internal reliability as measured by the coefficient alpha is equally high in both cases. However, it appears that the SFSS is measuring something different when the reference period is short versus long, which would suggest that enough respondents actually pay attention to the reference period for it to be of relevance.

If my conclusions are correct, it will be important to investigate what are the sources for the differences in the two-week and the longer versions. Are these differences the same for all types of items or are there differences among the items as the model in Figure 1 would suggest? Are the differences more or less pronounced based on certain respondent characteristics (e.g., age)? I conducted several explorative analyses to shed some light on the answers to these questions and point out some directions for future research.

Differences in Items

In order to explore if the differences between the reference and the comparison group differ across items, and if so, whether there may be some systematic trend present, I created two tables. Table 14 in Appendix B lists the frequencies for each answer choice (i.e., *Never, Hardly Ever, Sometimes, Often* and *Very Often*) item by item for each group. Frequencies that differ by more than 7% are highlighted[14]. Table 15 (youth) and Table 16 (caregiver) in Appendix B present the differences in means as well as the results of traditional tests for significance using Bonferroni adjustments to control for the 33 simultaneous tests. Items that showed significant differences are highlighted. Here I will discuss the most interesting findings of this explorative work.

The first noticeable thing is that there are many more items that show differences between the short and longer reference periods in the caregiver sample than in the youth sample. However, since this refers more to differences in respondent characteristics I will discuss this in the next section.

The next thing that I found interesting is that those items which demonstrate significant mean differences and also have the most pronounced mean differences (< -.4) in the caregiver sample are almost exclusively items that according to the Diagnostic and Statistical Manual of Mental Disorders – Fourth Edition (DSM-IV, 1994), describe youth with conduct and oppositional defiant disorder (except for one item in each sample that would fall under hyperactivity and impulsivity). As in previous cases, the longer reference period results in higher ratings on each of these items. There could be several explanations for this. For one, it could be that youths who have this disorder are more

[14] Because this work is explorative I used a 7% criterion rather than the 10% used for testing the differences in proportions in hypothesis 2.

likely to admit their defiant behaviors if asked about the last three or six months than during a more recent time period. This would be an example of differential editing for social desirability during the final response process. However, that would not necessarily explain why there are differences in the caregiver sample as well. It could also be that there are substantive differences. That is, the likelihood that these types of behaviors occur in a period of two weeks is much lower than the likelihood that they occur in the longer periods of three or six months. However, there are other items for which a substantive difference seems much more intuitive such as using drugs for non-medical purposes and being involved in fights, however responses to these items had no significant differences in either sample. Because of the limitations of the current dataset I will not be able to determine a definitive answer to these questions. However, it clearly shows a need for pursuing a better understanding of the response process for these types of clinical measures.

Reviewing the frequencies in Table 14, one notices another interesting phenomenon in the data. For the youth sample, 14 items differ on the answer category *never* out of those 16 items which show a difference greater than 7%. In the caregiver sample, 23 items have answer categories for which the frequencies differ by more than 7% and for 18 of those they differ on the answer category *never*. It must be noted, though, that in the caregiver sample the items also differ in regard to other answer categories. But, the trend here is clearly discernable even without a formal test. In the introductory chapter I suggested that if there is a high level of ambiguity in the question (which probably applies to most of the questions in the SFSS) than we are most likely to find differences in the *never* category because it is not as vague and, thus, relative and

subjective as the other answer choices. The trend found in the data clearly seems to support this. If the model presented in Figure 1 is correct, however, this would imply that the respondents actually paid attention to the reference period. Furthermore, this points out the problem of using vague quantifiers as answer options. Even though it is very likely that the respondents recall different numbers of events (e.g., the number of times the youth felt worthless) when asked about the past two weeks or six months, their final response on the questionnaire may be the same for both types of response unless they are very sure that it never happened within the last two weeks. To me, this is related to the points raised by Blanton and Jaccard (2006) about the arbitrariness of many psychological scales. What does it mean to the comparability of results obtained with the same scale but different reference periods? As discussed earlier, several scale developers suggest using their scale with whatever reference period seems appropriate to the user. Before I discuss this further, though, let us consider differences by respondent characteristics.

Differences by Respondent Characteristics

One source of differences based on respondent characteristics seems clear if one studies Tables 14, 15, and 16 discussed in the previous section. There are many more items with significant mean difference (22 compared to 6) and differences in the answer choice frequencies (23 compared to 16) in the caregiver sample than there are in the youth sample. Two possible reasons for these differences come to mind. First, it could be that as proxy reporters the caregivers have different knowledge or memory of the events to be recalled. It is noticeable, for example, that while there are no mean differences for any of the items asking about experienced emotions in the youth sample, several of those

items show mean differences in the caregiver sample. Experienced emotions of another person are often not observable. Thus, for many of the items asking about emotions caregivers will have to guess. If these caregivers also believe that their child has changed, then, according to my application of the review by Robinson and Clore (2002) presented in the introduction, it is possible that they respond to a six month reference period differently than they would to a two-week version. If this would be indeed the reason for the differences, one would expect greater differences for those youths who have been in treatment longer because the expectation that they changed is more likely for these youths. Below I will present some preliminary evidence that this may in fact be the case.

The other possible reason for differences between youths and caregivers could be their difference in age. Response rates to surveys, for example, are often lower for youth than they are for adults (Watson and Wooden, 2006). This may suggest that youths are less motivated to participate in the survey process, which could affect how they react to completing mental health outcome questionnaires. That is, they may pay less attention to the details of the instructions (including the reference period) and also spend less effort on answering the questions accurately. In addition, younger youths have less experience with completing questionnaires, may have a harder time reading, and are less developed in their cognitive ability than the older youth. It is reasonable to assume that all of these reasons could lead to less pronounced differences between the reference group and the comparison group in the youth sample compared to the caregiver sample.

I conducted additional explorative analyses to investigate some of these ideas in regard to the difference in treatment length and age. Since there was no reason for me to expect that age would matter for the caregivers I only explored this variable in the youth

sample. In addition to the youths' age and length of treatment, I also investigated whether the difference in the youth's externalizing and internalizing sub-scale score mattered. For this purpose I created a new variable that is based on the simple difference between the externalizing and internalizing sub-scale scores based on the original SFSS scoring (range 32 - 92 for youths and 42 - 94 for caregivers). Thus, a positive value on this variable represents a higher score on the externalizing scale than on the internalizing scale while a negative score would represent the opposite. I investigated the potential moderating effects of these covariates in two ways: 1) I created cross-tables of the experimental condition by several levels of the covariate; 2) I included an interaction term of the treatment variable with each of these covariates into the mixed model used for generating the estimates for mean differences and their standard error in the test of hypothesis 1. None of the interaction terms were significant using the second method. Since the detection of interaction effects often requires large sample for sufficient statistical power, some of the non-significant result could be simply due to lack of power. Future investigation will have to determine this. Thus, I will concentrate my discussion on the cross-tables presented in Table 17 in Appendix B. I encourage the reader to consider these preliminary findings with caution since they are based on explorative work and do not involve planned and formal hypothesis tests. I included them here mainly to guide future research.

In regard to length of treatment, there is no clear trend discernable in the caregiver sample. In the youth sample, however, it is interesting to note that the difference between the reference and the comparison group is positive (i.e., the two-week version results in higher severity scores on average) in the first three months of treatment, but then

becomes negative by quite a margin after that. Reviewing the actual mean scores in each group one can see that this trend is mainly due to the changes in the two-week version. That is, using the two-week version the average severity level is high for youths who are early in their treatment and consistently gets lower the longer they have been in treatment. This is what one would expect. However, this same trend cannot be observed for the three- and six-month version. I leave it to the reader to decide what the implications of this are for the validity of using a scale with a long reference period such as three or six months.

In regard to age, there seems to be a difference between the younger youths (age 11 and 12) and older youths. Interestingly, the difference between the reference group and the control group is more pronounced among the younger youths. Maybe the younger youths are more motivated to answer the questions accurately than the older teenagers. Future research should look into this.

Another variable I explored is the difference between the score on the externalizing sub-scale and the internalizing-subscale. There is no clear trend observable in the youth sample in regard to this variable. The situation is different in the caregiver sample, however. The magnitude of the difference between the reference group and the comparison group is almost one standard deviation if the internalizing score is significantly higher (by one standard deviation) than the externalizing score. In the case were both scores are about equal, the difference is still quite pronounced (-.62). In both of these cases the scores on the two-week version are lower. However, in cases where the externalizing scores are significantly higher (by one standard deviation) than the internalizing scores, the differences between the reference and the comparison group

become negligible. It is interesting to note that for the two-week version the mean scores are about half a standard deviation above zero (i.e., higher severity than average) when the externalizing score is higher, but about half a standard deviation below zero (i.e., lower severity than average) if the two sub-scale scores are equal or the internalizing one is substantially higher. The same is not true for the three- and six-month versions. In fact, the mean is zero (i.e., average) if both sub-scale scores are about equal and about .4 if there are differences in either direction. I am not sure what this implies other than that the two versions seem to work differently. A more thorough and definitive testing of this phenomenon needs to be conducted to ensure that this is not an artifact of the current data.

Besides the variables discussed above, I also explored the familiarity of caregiver to the youth. It is reasonable to assume that foster parents who only recently became the caregiver of a youth will have a hard time answering questions with a three- or six-month reference period. But, would one expect the caregiver to rate youths that they do not know well higher on the three- or six-months version or lower in severity compared to the two-week version? The data presented in Table 17 would suggest that caregivers who indicated they know the youth *Not to well* or *Fairly well* rate the youth significantly higher by quite a margin if the longer reference period is used. While these numbers are based on a very small sample (12 caregivers said they know the youth *Not to well* and 33 said they know the youth *Fairly well*, and 5 in this sub-sample said they knew the youth *Very well*[15]) the differences are striking. The difference for the group who said they know the youth *Not to well* is almost two standard deviations (-1.88) while it is only slightly

[15] Only caregivers who indicated that they know the youth less than a year were asked to answer this question about familiarity with the youth.

lower for those who said they know the youth *Fairly well* (-1.73). This is very different from the average of the whole sample and raises concerns about using foster parents who have been with the child for only a little time as informants about the youths' severity and functioning. Future research will have to confirm this preliminary finding and investigate the reasons for these large differences.

Limitations

This current study has several limitations. First, the study is cross-sectional and does not allow for a comparison within youths and caregivers. Thus, I was only able to make inferences at the group level and how the scores compare across similar individuals. I cannot provide any insight into how these different reference periods would affect the inferences drawn from the study of individual growth trajectories over the course of treatment. Second, I do not have information about the actual response process of the youths and caregivers. All that is available to me are their final answer choices. While the explorative analyses presented in this discussion may provide some first ideas of what some of the sources of the differences could be, the current study does not provide any definitive answer in regard to the sources of the differences. It is only a first step in highlighting the fact that there are differences that are worth exploring. Clearly, the limited sample size for the test of hypothesis 3 is a limitation as well. There was not sufficient power for conducting the test using this smaller sub-sample. The great advantage of this hypothesis test would have been that it tests the scores of the same youths and caregivers on similar scales (the SFSS and CBCL/YSR). In one group the two scales would have matched reference periods (six months and six months) while in the other they would have very different reference periods (two weeks and six months).

Thus, this would be a simultaneous test of within and between group differences and, thus, may overcome some of the limitations of the cross-sectional design. Finally, the questionnaires were administered by the clinician during a treatment session. Even though the youths and caregivers were instructed to complete the questionnaires by themselves and confidentiality was promised to the youths and caregivers, we have no knowledge to what degree this way of collecting the data affected the answer choices of the youths. If indeed social desirability is affected differently if different reference periods are used, then the clinicians' involvement in administration of the measure could have introduced a methodological artifact. However, it is important to note that in most practical settings data is collected this way. That is, if there is a methodological artifact, it will be present in the actual application of this scale as well.

Conclusions and Suggestions for Future Research

In this dissertation I have stressed the need to better understand the methodological implications of relying on self and proxy-reports in the context of youth mental health outcomes measures. I focused my investigation on the impact of using a short vs. a long reference period on the response of youths and caregivers. I held methodological factors other than the reference period constant by using just one measure, the SFSS). Based on the review of the literature I developed a model that led me to believe that, at the group level, I should not expect any meaningful differences between the reference and the comparison group in the current study. I defined meaningful differences in regard to four aspects: means, proportions of youths in the high severity range, correlation with a validity scale, and internal reliability assessed by the

coefficient alpha. Because I expected no differences I used methods to test for the equivalence of two groups in regard to certain population parameters.

Based on the findings of these tests, I conclude that there are meaningful differences caused by the use of short vs. longer reference periods. These differences did not affect the reliability estimates of the SFSS which are the same in both the reference and the comparison groups. The differences that surfaced affect the determination of the level of youths' symptom severity and functioning. Because this implies that my original hypotheses about no differences were disconfirmed I felt compelled to investigate these differences a little bit further. In the discussion section I presented some highlights of this explorative work in regard to differences by items as well as respondent types.

What does all this mean for the use of these types of youth mental health outcome scales? For example, if the magnitude and direction of the difference between the two-week and the longer version really differs by the length the youths have been in treatment, are we assessing the same effect of treatment if different references periods are used? In most cases that is not likely. If caregivers, who do not know the youth well, are so strongly affected by the reference period, should they even be asked to complete the questionnaire? What implications do these differences have for a practitioner who assessed a child at intake with a six-month version and then uses a two-week version from there on? It could be that they observe a decrease in severity that is simply due to the different reference period. This methodological issue becomes a practical one when treatments are being evaluated with self-report outcome measures such as the SFSS. With the knowledge I gained through this empirical study I would recommend using only one version, that is, the two-week version, throughout a study or the process of treatment

outcome monitoring. However, the need for future research in this area should be clear. Thus, I will end this dissertation by proposing three possible directions for future research.

First, in order to better understand the response process in regard to different item characteristics and in order to confirm some of the preliminary findings of the explorative work, I propose to categorize the items of the SFSS by characteristics that could explain difference according to the model presented in Figure 1. These categories would include, but not necessarily be limited to, experienced emotion vs. behavioral event, saliency, ambiguity, and social desirability. I have begun some pilot work in developing a rating scale in regard to these characteristics that several clinical experts will use in rating each of the items of the SFSS. These ratings can then be used to formally test whether there are significant differences or not (i.e., equivalence) in regard to the differences between the reference group and the comparison group completing measures with 2-week vs. three- and six-month reference periods respectively, based on these item characteristics.

Second, I propose using cognitive interviews to observe the actual response process. In cognitive interviews, respondents report either concurrently (think-aloud) or right after the completion of the questionnaire what their thoughts were when generating answers to the questions. This could shed some light on whether the youths and caregivers actually recall events and if so, if recall differs according to the length of the reference period. It could also provide information on how they map their answer to the provided answer categories. All of this would help us determine to what degree difference are substantive and to what degree they are due to the way questions are asked and questionnaires are designed.

Third, I would recommend studying these effects in the context of a longitudinal, randomized experiment in order to investigate the question of different reference periods at the individual level over time.

A. SAS CODE FOR MIXED MODEL

```
*************************************************************************
Mixed Model for obtaining estimates for mean difference and standard
error for H1 a)
*************************************************************************;

libname wkd "C:\dissertation\data";
libname library  "C:\dissertation\data";

ods rtf file = "C:\dissertation\results\h1a_mixedmodelestimates.rtf";

* mixed model with REPERIOD as fixed and OFFICE as random effect;

proc mixed data = wkd.finalyouthdata COVTEST;
      title 'Mixed model with REPERIOD as fixed and OFFICE as random
            effect';
      ID Y1UNIQUEID;
      class OFFICE REPERIOD;
      model  Y_MEASURE=REPERIOD /solution;
      random OFFICE /solution;
      lsmeans REPERIOD / pdiff cl;
      make 'CovParms' out = COVparms ;
run;
quit;
```

B. RESULTS ITEM AND RESPONDENT ANALYSIS

Table 14: Frequencies of Answer Choices by Item

		Youths			Caregivers		
		2 W	3/6 M		2 W	3/6 M	
Item	**Choice**	**%**	**%**	**Diff.**	**%**	**%**	**Diff.**
1) Throw things when mad	Never	51.28	36.68	*14.6*	41.67	26.55	*15.12*
	Hardy ever	15.38	15.28	0.1	21.53	14.69	6.84
	Sometimes	24.1	29.69	-5.59	22.92	38.98	*-16.06*
	Often	4.62	10.04	-5.42	9.03	11.86	-2.83
	Very often	4.62	8.3	-3.68	4.86	7.91	-3.05
2) Eat a lot more or less	Never	35.05	27.75	*7.3*	27.4	18.13	*9.27*
	Hardy ever	9.79	11.45	-1.66	17.12	14.62	2.5
	Sometimes	28.35	33.48	-5.13	32.88	36.84	-3.96
	Often	11.86	16.3	-4.44	13.7	17.54	-3.84
	Very often	14.95	11.01	3.94	8.9	12.87	-3.97
3) Feel unhappy or sad	Never	14.43	12.28	2.15	8.9	3.41	5.49
	Hardy ever	16.49	17.98	-1.49	16.44	11.36	5.08
	Sometimes	41.75	36.4	5.35	54.11	46.59	*7.52*
	Often	10.82	21.93	*-11.11*	13.7	26.7	*-13*
	Very often	16.49	11.4	5.09	6.85	11.93	-5.08
4) Get into trouble	Never	17.71	10.13	*7.58*	15.75	7.47	*8.28*
	Hardy ever	18.75	17.18	1.57	21.23	11.49	*9.74*
	Sometimes	38.02	36.12	1.9	35.62	41.95	-6.33
	Often	15.63	18.06	-2.43	16.44	24.14	*-7.7*
	Very often	9.9	18.5	*-8.6*	10.96	14.94	-3.98
5) Have little or no energy	Never	40	34.21	5.79	30.82	19.43	*11.39*
	Hardy ever	20.51	20.18	0.33	24.66	29.71	-5.05
	Sometimes	25.13	24.12	1.01	28.08	28.57	-0.49
	Often	7.69	10.09	-2.4	12.33	11.43	0.9
	Very often	6.67	11.4	-4.73	4.11	10.86	-6.75
6) Disobey adults	Never	17.44	13.16	4.28	11.72	5.11	6.61
	Hardy ever	17.95	17.54	0.41	18.62	13.64	4.98
	Sometimes	37.95	42.98	-5.03	33.79	38.07	-4.28
	Often	13.33	13.16	0.17	17.24	21.02	-3.78
	Very often	13.33	13.16	0.17	18.62	22.16	-3.54
7) Interrupt others	Never	21.65	21.83	-0.18	13.7	6.82	6.88
	Hardy ever	21.65	23.14	-1.49	14.38	13.64	0.74
	Sometimes	36.6	35.81	0.79	34.93	30.11	4.82
	Often	10.82	10.48	0.34	17.81	24.43	-6.62
	Very often	9.28	8.73	0.55	19.18	25	-5.82

Item	Choice	Youths 2 W %	Youths 3/6 M %	Youths Diff.	Caregivers 2 W %	Caregivers 3/6 M %	Caregivers Diff.
8) Lie to get things	Never	38.14	31.44	6.7	15.28	9.09	6.19
	Hardy ever	22.16	27.51	-5.35	21.53	18.75	2.78
	Sometimes	24.74	27.07	-2.33	34.72	31.82	2.9
	Often	9.79	6.55	3.24	15.97	20.45	-4.48
	Very often	5.15	7.42	-2.27	12.5	19.89	*-7.39*
9) Hard time c temper	Never	22.68	17.98	4.7	12.5	6.86	5.64
	Hardy ever	13.92	16.67	-2.75	18.75	10.29	*8.46*
	Sometimes	33.51	33.77	-0.26	37.5	34.86	2.64
	Often	13.4	12.72	0.68	11.81	23.43	*-11.62*
	Very often	16.49	18.86	-2.37	19.44	24.57	-5.13
10) Use drugs non-medical	Never	84.02	78.41	5.61	82.64	85.47	-2.83
	Hardy ever	6.19	6.61	-0.42	7.64	5.23	2.41
	Sometimes	5.67	8.81	-3.14	8.33	4.07	4.26
	Often	1.55	4.41	-2.86	0.69	4.07	-3.38
	Very often	2.58	1.76	0.82	0.69	1.16	-0.47
11) Worry a lot	Never	20.31	20.09	0.22	9.59	7.43	2.16
	Hardy ever	14.06	15.63	-1.57	16.44	13.14	3.3
	Sometimes	31.25	25.45	5.8	47.26	45.71	1.55
	Often	17.71	16.52	1.19	15.75	20.57	-4.82
	Very often	16.67	22.32	-5.65	10.96	13.14	-2.18
12) Getting along w/ family	Never	30.93	20.52	*10.41*	15.07	7.95	*7.12*
	Hardy ever	17.53	19.21	-1.68	25.34	14.2	*11.14*
	Sometimes	29.38	30.13	-0.75	30.82	39.2	*-8.38*
	Often	10.31	15.72	-5.41	19.18	21.02	-1.84
	Very often	11.86	14.41	-2.55	9.59	17.61	*-8.02*
13) Threaten or bully others	Never	63.08	53.95	*9.13*	35.42	28.74	6.68
	Hardy ever	16.41	17.98	-1.57	25.69	17.24	*8.45*
	Sometimes	14.36	17.11	-2.75	20.83	26.44	-5.61
	Often	3.08	3.95	-0.87	13.19	18.97	-5.78
	Very often	3.08	7.02	-3.94	4.86	8.62	-3.76
14) Think about hurting yourself	Never	74.07	67.26	6.81	61.11	51.43	*9.68*
	Hardy ever	8.47	11.5	-3.03	22.22	18.86	3.36
	Sometimes	10.05	13.72	-3.67	13.19	24	*-10.81*
	Often	3.17	4.42	-1.25	1.39	2.86	-1.47
	Very often	4.23	3.1	1.13	2.08	2.86	-0.78
15) Feel worthless	Never	60.51	51.32	*9.19*	40.69	28.41	*12.28*
	Hardy ever	13.85	15.79	-1.94	19.31	24.43	-5.12
	Sometimes	14.87	16.67	-1.8	31.72	26.7	5.02
	Often	6.67	10.96	-4.29	6.9	14.2	*-7.3*
	Very often	4.1	5.26	-1.16	1.38	6.25	-4.87

Item	Choice	Youths 2 W %	Youths 3/6 M %	Youths Diff.	Caregivers 2 W %	Caregivers 3/6 M %	Caregivers Diff.
16) Drink alcohol	Never	86.67	78.51	*8.16*	83.56	80.11	3.45
	Hardy ever	3.59	10.53	-6.94	8.22	8.52	-0.3
	Sometimes	6.15	7.89	-1.74	6.85	7.39	-0.54
	Often	2.56	2.19	0.37	0.68	2.84	-2.16
	Very often	1.03	0.88	0.15	0.68	1.14	-0.46
17) Hard time having fun	Never	51.03	50.44	0.59	35.86	25.14	*10.72*
	Hardy ever	16.49	11.4	5.09	26.21	24.57	1.64
	Sometimes	20.1	25.88	-5.78	28.97	35.43	-6.46
	Often	6.19	5.7	0.49	6.21	10.86	-4.65
	Very often	6.19	6.58	-0.39	2.76	4	-1.24
18) Afraid others would laugh	Never	57.95	53.07	4.88	31.03	25.99	5.04
	Hardy ever	14.87	16.23	-1.36	20.69	15.25	5.44
	Sometimes	16.92	17.98	-1.06	34.48	32.2	2.28
	Often	6.67	5.7	0.97	7.59	16.38	*-8.79*
	Very often	3.59	7.02	-3.43	6.21	10.17	-3.96
19) Hard time waiting turn	Never	45.88	38.22	*7.66*	23.29	19.77	3.52
	Hardy ever	19.07	23.11	-4.04	21.92	22.09	-0.17
	Sometimes	21.65	21.33	0.32	30.14	30.81	-0.67
	Often	5.15	6.22	-1.07	10.96	16.28	-5.32
	Very often	8.25	11.11	-2.86	13.7	11.05	2.65
20) Sleep a lot more	Never	38.86	37.99	0.87	34.25	22.03	*12.22*
	Hardy ever	21.24	14.41	6.83	23.29	32.2	*-8.91*
	Sometimes	19.17	25.76	-6.59	26.71	31.64	-4.93
	Often	8.81	10.48	-1.67	10.96	8.47	2.49
	Very often	11.92	11.35	0.57	4.79	5.65	-0.86
21) Hang with kids in trouble	Never	41.67	32.02	*9.65*	42.07	32.18	*9.89*
	Hardy ever	17.19	21.49	-4.3	21.38	17.24	4.14
	Sometimes	29.69	26.32	3.37	22.76	28.74	-5.98
	Often	7.29	10.53	-3.24	6.21	11.49	-5.28
	Very often	4.17	9.65	-5.48	7.59	10.34	-2.75
22) Feel nervous around people	Never	40.41	34.06	6.35	33.56	22.6	*10.96*
	Hardy ever	20.21	18.78	1.43	23.29	25.99	-2.7
	Sometimes	23.32	26.2	-2.88	31.51	34.46	-2.95
	Often	7.25	11.35	-4.1	8.22	11.3	-3.08
	Very often	8.81	9.61	-0.8	3.42	5.65	-2.23
23) Hard time paying attention	Never	27.98	22.81	5.17	15.75	8.47	*7.28*
	Hardy ever	19.17	13.16	6.01	16.44	8.47	*7.97*
	Sometimes	26.94	32.89	-5.95	37.67	36.16	1.51
	Often	11.4	11.84	-0.44	17.12	23.73	-6.61
	Very often	14.51	19.3	-4.79	13.01	23.16	*-10.15*

Item	Choice	Youths			Caregivers		
		2 W	3/6 M		2 W	3/6 M	
		%	%	Diff.	%	%	Diff.
24) Fights	Never	38.22	31.44	6.78	21.92	17.05	4.87
	Hardy ever	20.42	19.65	0.77	20.55	13.64	6.91
	Sometimes	27.75	26.2	1.55	29.45	31.25	-1.8
	Often	4.71	11.79	*-7.08*	14.38	19.89	-5.51
	Very often	8.9	10.92	-2.02	13.7	18.18	-4.48
25) Lose things you need	Never	31.09	19.21	*11.88*	20	10.86	*9.14*
	Hardy ever	21.76	22.27	-0.51	23.45	18.86	4.59
	Sometimes	25.39	35.37	*-9.98*	33.1	38.29	-5.19
	Often	12.44	10.48	1.96	11.03	19.43	*-8.4*
	Very often	9.33	12.66	-3.33	12.41	12.57	-0.16
26) Hard time sitting still	Never	38.02	28.07	*9.95*	19.58	16.09	3.49
	Hardy ever	15.1	15.79	-0.69	20.28	16.67	3.61
	Sometimes	21.35	27.19	-5.84	27.27	33.33	-6.06
	Often	10.94	11.84	-0.9	17.48	16.09	1.39
	Very often	14.58	17.11	-2.53	15.38	17.82	-2.44
27) Hard time sleeping	Never	53.09	44.3	*8.79*	41.1	29.55	*11.55*
	Hardy ever	12.37	17.11	-4.74	19.86	26.7	-6.84
	Sometimes	15.98	18.42	-2.44	26.03	27.27	-1.24
	Often	8.76	10.09	-1.33	8.9	9.09	-0.19
	Very often	9.79	10.09	-0.3	4.11	7.39	-3.28
28) Feel tense	Never	46.6	40.79	5.81	17.12	15.34	1.78
	Hardy ever	16.23	17.11	-0.88	28.08	23.3	4.78
	Sometimes	19.37	23.25	-3.88	41.1	39.77	1.33
	Often	10.99	7.89	3.1	8.22	13.07	-4.85
	Very often	6.81	10.96	-4.15	5.48	8.52	-3.04
29) Cry easily	Never	44.04	43.17	0.87	36.99	28.98	*8.01*
	Hardy ever	17.62	19.82	-2.2	30.14	19.89	*10.25*
	Sometimes	15.54	21.15	-5.61	18.49	31.82	*-13.33*
	Often	7.77	6.17	1.6	7.53	11.36	-3.83
	Very often	15.03	9.69	5.34	6.85	7.95	-1.1
30) Annoy other people	Never	35.38	35.53	-0.15	19.18	13.71	5.47
	Hardy ever	24.1	17.54	6.56	14.38	12.57	1.81
	Sometimes	26.15	27.19	-1.04	30.14	36.57	-6.43
	Often	7.18	9.21	-2.03	21.23	16.57	4.66
	Very often	7.18	10.53	-3.35	15.07	20.57	-5.5
31) Argue with adults	Never	23.32	19.38	3.94	20.55	11.93	*8.62*
	Hardy ever	17.62	18.06	-0.44	15.07	7.95	*7.12*
	Sometimes	34.2	32.6	1.6	26.71	33.52	-6.81
	Often	14.51	17.18	-2.67	15.07	23.3	*-8.23*
	Very often	10.36	12.78	-2.42	22.6	23.3	-0.7

Item	Choice	Youths			Caregivers		
		2 W	3/6 M		2 W	3/6 M	
		%	%	Diff.	%	%	Diff.
32) Don't have any friends	Never	67.54	57.08	*10.46*	39.04	28.41	*10.63*
	Hardy ever	6.28	10.62	-4.34	22.6	27.84	-5.24
	Sometimes	15.71	17.7	-1.99	26.03	26.14	-0.11
	Often	5.24	7.08	-1.84	6.16	8.52	-2.36
	Very often	5.24	7.52	-2.28	6.16	9.09	-2.93
33) Too scared to ask in class	Never	64.92	55.46	*9.46*	48.28	36.31	*11.97*
	Hardy ever	4.71	13.1	*-8.39*	20	25.6	-5.6
	Sometimes	18.32	15.28	3.04	22.76	22.62	0.14
	Often	3.66	5.24	-1.58	6.21	7.74	-1.53
	Very often	8.38	10.92	-2.54	2.76	7.74	-4.98

Note. Differences > 7% are shaded

Table 15: Item Mean Differences - Youth

		2 Weeks			3/6 Months			Mean Difference		
	Item	N	Mean	SD	N	Mean	SD	Diff.	SE	t
1	Throw things when mad	195	1.96	1.17	229	2.38	1.29	-0.42	0.12	*-3.49*
2	Eat a lot more or less	194	2.62	1.44	227	2.71	1.32	-0.10	0.13	-0.70
3	Feel unhappy or sad	194	2.98	1.23	228	3.02	1.16	-0.04	0.12	-0.32
4	Get into trouble	192	2.81	1.19	227	3.18	1.21	-0.36	0.12	*-3.08*
5	Have little or no energy	195	2.21	1.23	228	2.44	1.35	-0.24	0.13	-1.88
6	Disobey adults	195	2.87	1.24	228	2.96	1.17	-0.08	0.12	-0.72
7	Interrupt others	194	2.64	1.20	229	2.61	1.19	0.03	0.12	0.28
8	Lie to get things	194	2.22	1.20	229	2.31	1.19	-0.09	0.12	-0.80
9	Hard time c temper	194	2.87	1.35	228	2.98	1.33	-0.11	0.13	-0.82
10	Use drugs non-medical	194	1.32	0.86	227	1.44	0.95	-0.12	0.09	-1.35
11	Worry a lot	192	2.96	1.34	224	3.05	1.42	-0.09	0.14	-0.66
12	Getting along w/ family	194	2.55	1.34	229	2.84	1.32	-0.30	0.13	*-2.29*
13	Threaten or bully others	195	1.67	1.03	228	1.92	1.22	-0.25	0.11	*-2.29*
14	Think about hurting self	189	1.55	1.07	226	1.65	1.07	-0.10	0.11	-0.91
15	Feel worthless	195	1.80	1.16	228	2.03	1.26	-0.23	0.12	-1.94
16	Drink alcohol	195	1.28	0.78	228	1.36	0.79	-0.09	0.08	-1.14
17	Hard time having fun	194	2.00	1.23	228	2.07	1.26	-0.07	0.12	-0.54
18	Afraid others would laugh	195	1.83	1.15	228	1.97	1.26	-0.14	0.12	-1.21
19	Hard time waiting turn	194	2.11	1.27	225	2.29	1.33	-0.18	0.13	-1.41
20	Sleep a lot more	193	2.34	1.38	229	2.43	1.38	-0.09	0.13	-0.68
21	Hang with kids in trouble	192	2.15	1.17	228	2.44	1.30	-0.29	0.12	*-2.40*
22	Feel nervous around people	193	2.24	1.29	229	2.44	1.32	-0.20	0.13	-1.55
23	Hard time paying attention	193	2.65	1.38	228	2.92	1.39	-0.26	0.14	-1.95
24	Fights	191	2.26	1.26	229	2.51	1.33	-0.25	0.13	-2.00
25	Lose things you need	193	2.47	1.30	229	2.75	1.24	-0.28	0.12	*-2.25*
26	Hard time sitting still	192	2.49	1.45	228	2.74	1.42	-0.25	0.14	-1.79
27	Hard time sleeping	194	2.10	1.39	228	2.25	1.37	-0.15	0.13	-1.10
28	Feel tense	191	2.15	1.30	228	2.31	1.36	-0.16	0.13	-1.22
29	Cry easily	193	2.32	1.47	227	2.19	1.32	0.13	0.14	0.94
30	Annoy other people	195	2.27	1.22	228	2.42	1.33	-0.15	0.13	-1.20
31	Argue with adults	193	2.71	1.26	227	2.86	1.28	-0.15	0.12	-1.20
32	Don't have any friends	191	1.74	1.21	226	1.97	1.31	-0.23	0.12	-1.85
33	Too scared to ask in class	191	1.86	1.31	229	2.03	1.38	-0.17	0.13	-1.30

Note. Significant differences at the $\alpha = .05$ level (with Bonferroni adjustment) are shaded in grey

Table 16: Item Mean Difference - Caregiver

		2 Weeks			3/6 Months			Mean Difference		
	Item	N	Mean	SD	N	Mean	SD	Diff.	SE	t
1	Throw things when mad	144	2.14	1.20	177	2.60	1.22	-0.46	0.14	*-3.38*
2	Eat a lot more or less	146	2.60	1.27	171	2.92	1.25	-0.33	0.14	*-2.31*
3	Feel unhappy or sad	146	2.93	0.97	176	3.32	0.95	-0.39	0.11	*-3.67*
4	Get into trouble	146	2.86	1.20	174	3.28	1.09	-0.42	0.13	*-3.28*
5	Have little or no energy	146	2.34	1.16	175	2.65	1.23	-0.30	0.13	*-2.26*
6	Disobey adults	145	3.12	1.25	176	3.41	1.13	-0.29	0.13	*-2.19*
7	Interrupt others	146	3.14	1.28	176	3.47	1.20	-0.33	0.14	*-2.37*
8	Lie to get things	144	2.89	1.22	176	3.23	1.23	-0.34	0.14	*-2.50*
9	Hard time c temper	144	3.07	1.26	175	3.49	1.17	-0.42	0.14	*-3.05*
10	Use drugs non-medical	144	1.29	0.71	172	1.30	0.82	-0.01	0.09	-0.12
11	Worry a lot	146	3.02	1.07	175	3.19	1.06	-0.17	0.12	-1.40
12	Getting along w/ family	146	2.83	1.19	176	3.26	1.15	-0.43	0.13	*-3.32*
13	Threaten or bully others	144	2.26	1.21	174	2.61	1.31	-0.35	0.14	*-2.46*
14	Think about hurting self	144	1.61	0.92	175	1.87	1.06	-0.26	0.11	*-2.30*
15	Feel worthless	145	2.09	1.06	176	2.45	1.22	-0.36	0.13	*-2.83*
16	Drink alcohol	146	1.27	0.68	176	1.36	0.83	-0.10	0.09	-1.13
17	Hard time having fun	145	2.14	1.06	175	2.44	1.10	-0.30	0.12	*-2.48*
18	Afraid others would laugh	145	2.37	1.18	177	2.69	1.30	-0.32	0.14	*-2.31*
19	Hard time waiting turn	146	2.70	1.32	172	2.77	1.25	-0.07	0.14	-0.48
20	Sleep a lot more	146	2.29	1.19	177	2.44	1.10	-0.15	0.13	-1.16
21	Hang with kids in trouble	145	2.16	1.25	174	2.51	1.32	-0.35	0.15	*-2.39*
22	Feel nervous around	146	2.25	1.11	177	2.51	1.13	-0.27	0.13	*-2.13*
23	Hard time paying	146	2.95	1.22	177	3.45	1.18	-0.49	0.13	*-3.68*
24	Fights	146	2.77	1.32	176	3.09	1.32	-0.31	0.15	-2.11
25	Lose things you need	145	2.72	1.26	175	3.04	1.15	-0.32	0.13	*-2.34*
26	Hard time sitting still	143	2.89	1.33	174	3.03	1.30	-0.14	0.15	-0.95
27	Hard time sleeping	146	2.15	1.18	176	2.38	1.21	-0.23	0.13	-1.72
28	Feel tense	146	2.57	1.04	176	2.76	1.13	-0.19	0.12	-1.58
29	Cry easily	146	2.17	1.21	176	2.49	1.24	-0.32	0.14	*-2.35*
30	Annoy other people	146	2.99	1.32	175	3.18	1.28	-0.19	0.15	-1.31
31	Argue with adults	146	3.04	1.43	176	3.38	1.26	-0.34	0.15	*-2.27*
32	Don't have any friends	146	2.18	1.20	176	2.42	1.24	-0.24	0.14	-1.77
33	Too scared to ask in class	145	1.95	1.10	168	2.25	1.24	-0.30	0.13	*-2.23*

Note. Significant differences at the $\alpha = .05$ level (with Bonferroni adjustment) are shaded in grey

Table 17: Potential Moderator Effects

	Youths			Caregivers		
Variable	2 W	3/6 M	Diff.	2 W	3/6 M	Diff.
Length of Treatment						
0-1 month	0.36	0.17	0.19	-0.16	0.14	-0.30
1-3 months	0.16	0.02	0.14	-0.40	0.18	-0.58
3-6 months	-0.27	0.11	-0.38	-0.16	0.00	-0.16
6-12 months	-0.32	0.26	-0.58	-0.41	0.12	-0.53
> 12 months	-0.40	-0.09	-0.31	-1.04	0.00	-1.04
Youth's Age						
Age 11-12	-0.05	0.40	-0.45			
Age 13-16	-0.16	0.03	-0.19			
Age 17-18	-0.15	0.02	-0.17			
External. vs. Internal.						
- 1 Stdev	0.33	0.63	-0.30	-0.59	0.38	-0.97
0	-0.21	0.03	-0.24	-0.64	-0.02	-0.62
+1 Stdev	-0.08	0.14	-0.22	0.51	0.42	0.09
Familiarity with child						
Not to well				-0.97	0.91	-1.88
Fairly well				-1.72	0.01	-1.73
Very well				-0.13	-0.49	0.36

REFERENCES

Achenbach, T. M. (1985). *Assessment and Taxonomy of Child and Adolescent Psychopathology: Volume 3. Development Clinical Psychology and Psychiatry.* Beverly Hills, California: Sage Publications.

Achenbach, T. M. (1991a). *Manual for the Child Behavior Checklist/4-18 and 1991 profile*. Burlington, VT: University of Vermont Department of Psychiatry.

Achenbach, T. M. (1991b). *Manual for the Youth Self-Report and 1991 profile.* Burlington, VT: University of Vermont Department of Psychiatry.

Achenbach, T. M. (1991c). *Manual of the Teacher's Report Form and 1991 profile.* Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M., McConoughy, S. H., & Howell, C**.** T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101,* 213-232.

Anderson, S. & Hauck, W.W. (1983). A new procedure for testing equivalence in comparative bioavailability trials. *Communications in Statistics - Theory and Methods, 12*, 2663-2692.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? In E.V. Smith, Jr. and R.M. Smith (Eds.). *Introduction to Rasch Measurement* (pp. 143-166). Maple Grove, MN: JAM Press.

Bachman, J. G., & O'Malley, P. M. (1981). When four months equal a year: Inconsistencies in student reports of drug use. *Public Opinion Quarterly, 45* (4), 536-548.

Baldwin, W. (2000). Information No One Else Knows: The Value of Self-Report. In A. A. Stone, J. S. Turkkan, C. S. Bachrach, J. B. Jobe, H. S. Kurztman, & V. S. Cain (Eds.), *The Science of Self-Report: Implications for Research and Practice* (pp. 229-255). Mahwah, NJ: Lawrence Erlbaum.

Barsalou, L.W. (1988). The content and organization of autobiographical memories. In U. Neisser and E. Winograd (Eds.), *Remembering Reconsidered: Ecological and Traditional Approaches to the Study of Memory* (pp. 193-243). Cambridge, England: Cambridge University.

Belli, R. F., Schwarz, N., Singer, E., & Talarico, J. (2000). Decomposition can harm the accuracy of behavioural frequency reports. *Applied Cognitive Psychology, 14* (4), 295-308.

Belson, W.A. (1981). *The design and understanding of survey questions*. Aldershot, U.K.: Gower.

Berger, L.R., and Hsu, J.C. (1996), Bioequivalence trials, intersection-union tests and equivalence confidence sets, *Statistical Science, 11,* 283-302

Bickman, L., Breda, C., Dew, S., Lambert E. W., Pinkard, T. J., Riemer, M., Vides de Andrade, A. R., Westlake, M. W. (Eds.) (2006). *Peabody Treatment Progress Battery Manual* [Electronic version]. Nashville, TN: Vanderbilt University. http://peabody.vanderbilt.edu/ptpb/

Bickman, L., Nurcombe, B., Townsend, C., Belle, M., Schut, J., and Karver, M. (1998/9). *Consumer Measurement Systems in Child and Adolescent Mental Health.* Canberra, ACT: Department of Health and Family Services.

Blair, E., & Burton, S. (1987). Cognitive processes used by survey respondents to answer behavioral frequency questions. *Journal of Consumer Research, 14* (2), 280-288.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61,* 27-41.

Bond, T. G. & Fox, C. M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Mahway, NJ: Earlbaum.

Bonett, D.G. (2003). Sample size requirements for comparing two alpha coefficients. *Applied Psychological Measurement, 27* (1), 72-74.

Brown, N.R., & Sinclair, R.C. (1997). *Estimating the number of lifetime sexual partners: men and women do it differently.* Paper presented at the 52nd Conference of the American Association for Public Opinion Research, May 15-18, Norfolk, VA.

Burlingame, G.M., Wells, M.G., Cox, J.C, Lambert, M.J., Latkowski, M., Ferre, R. (2005). Administration and Scoring Manual for the Youth Outcome Questionnaire. American Professional Credentialing Services LLC, Stevenson: MD.

Burton, S. & Blair, E. (1991). Task conditions, response formulation processes, and response accuracy for behavioral frequency questions in surveys. *Public Opinion Quarterly, 55*, 50-79

Clark, L. A. and D. Watson (1995). "Constructing Validity: Basic Issues in Objective Scale Development." *Psychological Assessment, 7* (3): 309-319.

Chu, A., Eisenhower, D., Hay, M., Morganstein, D., Neter. J., and Waksberg, J. (1992). Measuring the recall error in self-reported fishing and hunting activities. *Journal of Official Statistics, 8* (1), 19-39

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Conway, M. A. (1996). Autobiographical knowledge and autobiographical memories. In D.C. Rubin (Ed.), *Remembering Our Past: Studies in Autobiographical Memory* (pp. 67-93). Cambridge, England: Cambridge University.

Diagnostic and statistical manual of mental disorders (DSM-IV). (1994). Washington, D.C.: American Psychiatric Association.

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child and Adolescent Psychiatry, 40* (11), 1337-1345.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment 8,* 341–349.

Goodman, R. & Scott, S. (1999). Comparing the strengths and difficulties questionnaire and the Child Behavior Checklist: Is small beautiful? *Journal of Abnormal Child Psychology, 27* (1), 17-24.

Hays, W. L. (1994). *Statistics (5th ed.)*. Fort Worth, TX: Harcourt Brace College Publishers.

Henggeler, S.W., Rowland, M.D., Halliday-Boykins, C., Sheidow, A.J., Ward, D.M., Randall, J., Pickrel, S.G., Cunningham, P. B., Edwards, J. (2003). One-Year Follow-up of Multisystemic Therapy as an Alternative to the Hospitalization of Youths in Psychiatric Crisis. *Journal of the American Academy of Child & Adolescent Psychiatry, 42*, 543-551

Hodges, K. (1990). *Child and Adolescent Functional Assessment Scale (CAFAS)*. Ypsilanti, MI: Eastern Michigan University, Department of Psychology.

Igou, E. R., Bless, H., Schwarz, N. (2002). Making sense of standardized survey questions: The influence of reference periods and their repetition. *Communication Monographs, 69* (2), 179-187.

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology, 67* (3), 300-307.

Jacobson, N. S. & Truax, P. (1991). Clinical Significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59* (1), 12-19.

Jobe, J. B., & Herrmann, D. (1996). Implications of models of survey cognition for memory theory. In D. Herrmann, C. Johnson, C. McEvoy, C. Herzog & P. Hertel (Eds.), *Basic and applied memory research: Volume 2: Practical Applications* (pp. 193-205). Hillsdale, NJ: Erlbaum.

Kessler, R. C., Wittchen, H. U., Abelson, J., & Zhao, S. (2000). Methodological Issues in Assessing Psychiatric Disorders With Self-Reports. In A. A. Stone, J. S. Turkkan, C. S. Bachrach, J. B. Jobe, H. S. Kurztman, & V. S. Cain (Eds.), *The Science of Self-Report: Implications for Research and Practice* (pp. 229-255). Mahwah, NJ: Lawrence Erlbaum.

Kolodner, J.L. (1984). *Retrieval and Organizational Strategies in Conceptual Memory.* Hillsdale, NJ: Erlbaum.

Kraatz, M. (2006). *Testing Correlational Equivalence*. Paper presented at the Quantitative Methods Brown Bag, August, 28. Nashville, TN: Vanderbilt University.

Linacre, J.M. (1998). Detecting multidimensionality: Which residual data-types works best? *Journal of Outcome Measurement, 12*, 266-283.

Linacre, J.M. (2004). Estimation methods for Rasch measures. In E.V. Smith, Jr. and R.M. Smith (Eds.). *Introduction to Rasch Measurement* (pp. 25-47). Maple Grove, MN: JAM Press.

Littell, R.C., Milliken, G.A., Stroup, W.W., & Wolfinger, R.D. (1996). *SAS System for Mixed Models.* Cary, NC: SAS Institute Inc.

Loftus, E.F., Klinger, M.R., Smith, K.D., & Fiedler, J. (1990). A tale of two questions: Benefits of asking more than one question. *Public Opinion Quarterly, 54*, 330-345.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McFarland, C., & Ross, M. (1987). The relation between current impressions and memories of self and dating partners. *Personality and Social Psychology Bulletin, 13*, 228-238.

Meyer, G.J., Finn, S.E., Eyde, L., Kay, G.G., Moreland, K.L., Dies, R.R., Eisman, E.J., Kubiszyn, T.W., & Reed, G.M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56*, 128-165.

Milliken, G.A. & Johnson, D.E. (2002). *Analysis of Messy Data: Volume 3. Analysis of Covariance.* Boca Raton, FL: Chapman & Hall/CRC.

Muthen B.O. and Satorra A. (1995). Complex Sample Data in Structural Equation Modeling. *Sociological Methodology, 25*, 267-316.

Neter, J., & Waksberg, J. (1964). A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association, 59*, 17-55.

Ohio Department of Mental Health (2004). *The Ohio Mental Health Outcomes Consumer System: A Procedural Manual, (6th Ed.)*. Columbus, OH: Ohio Department of Mental Health.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72*, 407-418.

Parducci, A. (1974). Contextual effects: A range-frequency analysis. In E. Carterette & M. Friedman (Eds.), *Handbook of perception: Psychophysical judgment and measurement.* (Vol. II, pp. 127-141). New York: Academic Press.

Pastor, D.A. (2003). The Use of Multilevel Item Response Theory Modeling in Applied Research: *An Illustration. Applied Measurement in Education, 16* (3), 223-243

Phillips, K.F. (1990), "Power of the Two One-sided Tests Procedure in Bioequivalence," *Journal of Pharmacokinetics and Biopharmaceutics, 18*, 137-144.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests* (Rev. ed.). Chicago: University of Chicago Press.

Rips, L.J. (1995). The current status of research on concept combination. *Mind & Language, 10*, 72-104.

Robinson, M.D. & Clore, G.L. (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin, 128* (6), 934-960.

Rogers, J. L., Howard, K. L., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113* (3), 553-565.

Sander, J.E., Conrad, F.G., Mullin, P.A., & Herrmann, D. (1992). Cognitive modeling of the survey interview. *Proceedings of the Section on Survey Methods Research, American Statistical Association* (pp. 818-823).

Schaeffer, N.C. & Presser, S. (2003). The Science of Asking Questions. *Annual Review of Sociology, 29*, 65-88.

Schwarz, N., Strack, F., Müller, G., & Chassein, B. (1988). The range of response alternatives may determine the meaning of the question: Further evidence on informative functions of response alternatives. *Social Cognition, 6*, 107 - 117.

Seaman, M.A. & Serlin, R.C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods, 3* (4), 403-411.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2001). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Boston, MA: Houghton Mifflin.

Shiffman, S. (2000). Real-Time Self-Report of Momentary States in the Natural Environment: Computerized Ecological Momentary Assessment. In A.A. Stone, J.S. Turkkan, C.S. Bachrach, J.B. Jobe, H.S. Kurztman, & V.S. Cain (Eds.), *The Science of Self-Report: Implications for Research and Practice* (pp. 277-296). Mahwah, NJ: Lawrence Erlbaum.

Shum, M.S. & Rips, L.J. (1999). The respondent's confession: Autobiographical memory in the context of surveys. In M.G. Sirken, D.J. Herrmann, S. Schechter, N. Schwarz, J.M. Tanur, & R. Tourangeau (Eds.), *Cognition and Survey Research* (pp. 95-109). New York, NY: John Wiley & Sons, Inc.

Smith, A.F. (1991). Cognitive processes in long-term dietary recall. *Vital Health Statistics, 6* (4), 1–42.

Smith, E.V., (2004). Detecting and evaluation the impact of multidimensionality using item fit statistics and principal component analysis of residuals. In E.V. Smith, Jr. and R.M. Smith (Eds.). *Introduction to Rasch Measurement* (pp. 575-600). Maple Grove, MN: JAM Press.

Smith, R.M. (2004). Fit analysis in latent trait measurement models. In E.V. Smith, Jr. and R.M. Smith (Eds.). *Introduction to Rasch Measurement* (pp. 73-92). Maple Grove, MN: JAM Press.

Smith, R.M., Linacre, J.M., and Smith, E.V. (2003). Guidelines for Manuscripts. *Journal of Applied Measurement, 4*, 198-204.

Sperry, L., Brill, P. L., Howard, K. L., & Grissom, G. R. (1996). *Treatment outcomes in psychotherapy and psychiatric interventions.* Philadelphia: Brunner/Mazel, Inc.

Stegner, B., Bostrom, A., Greenfield, T. (1996). Equivalence testing for use in psychosocial and services research: An introduction with examples. *Evaluation and Program Planning*. *19* (3), 193-198.

Sudman, S., & Bradburn, N. (1973). Effects of time and memory factors on the response in surveys. *Journal of the American Statistical Association, 68*, 805-815.

Sudman, S. & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire design.* San Francisco: Jossey-Bass.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco, CA: Jossey-Bass.

Thomas, D.L. & Diener,E. (1990). Memory accuracy in the recall of emotions. *Journal of Personality and Social Psychology, 59* (2), 291-297.

Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling & Development, 80*, 64-71.

Thurstone, L.L. (1959). Attitudes can be measured. American Journal of Sociology, 1928, 33, 529-554, reprinted in The Measurement of Values, Chicago: University of Chicago Press.

Tourangeau, R. (1984). Cognitive science and survey methods. In T. Jabine, M. Straf, J. Tanur, R. Tourangeau (Eds.), Cognitive aspects of survey design: Building a bridge between disciplines (pp. 73-100). Washington, DC: National Academy Press.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response.* Cambridge, UK: Cambridge University.

Tulving, E. (1983). *Elements of episodic memory.* Oxford: Oxford University Press.

Watson, N. & Wooden, M. (2006). Identifying Factors Affecting Longitudinal Survey Response. Paper presented at the Methodology of Longitudinal Surveys International Conference, University of Essex, Colchester, UK, July 12-14. http://www.iser.essex.ac.uk/ulsc/mols2006/programme/data/papers/Watson.pdf

Wellek, S. (2003). *Testing Statistical Hypotheses of Equivalence.* Boca Raton, FL: Chapman & Hall/CRC.

Wells, M.G., Burlingame, G.M., Rose, P.M. (1999). Administration and Scoring Manual for the Self Report Version of the Youth Outcome Questionnaire. American Professional Credentialing Services LLC, Stevenson: MD.

Westlake, W.J. (1981). Response to T.B.L. Kirkwood: Bioequivalence testing – A need to rethink. *Biometrics, 37*, 589-594.

Winkielman, P., Knäuper, B., & Schwarz, N. (1998). Looking back at anger: Reference periods change the interpretation of emotion frequency questions. *Journal of Personality and Social Psychology, 75,* (3), 719-728.

Worthington, H. (2004). Methods for Pooling Results from Multi-center Studies. *Journal of Dental Research, 83*, 119-121.

Wright, B., & Linacre, M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370.

Wright, B., & Masters, G. (1982). *Rating scale analysis: Rasch measurement.* Chicago: MESA Press.

Wright, B. D. & Mok, M. M. C. (2004). An overview of the family of Rasch Measurement Models. In Smith, E.V. & Smith R. M. (Eds.). *Introduction to Rasch Measurement.* Maple Grove, MN: JAM Press.