AN INVESTIGATION OF THE PARENTING STRESS INDEX IN
THE CONTEXT OF GENERALIZABILITY THEORY

by

Jim D. Sharpnack

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Psychology

Approved:

UTAH STATE UNIVERSITY
Logan, Utah
1997
ABSTRACT

An Investigation of the Parenting Stress Index in the Context of Generalizability Theory

by

Jim D. Sharpnack, Master of Science
Utah State University, 1997

Major Professor: Dr. Susan L. Crowley
Department: Psychology

This present study examined the application of generalizability theory (GT) to the Parenting Stress Index (PSI) long and short forms for families having children with disabilities. The purpose of the study was to evaluate the dependability of parenting stress data scores gathered from families having children with disabilities. The data for the present study came from an extant data set collected by the Early Intervention Research Institute (EIRI; Contract #800-85-0173) at Utah State University. The EIRI studies represented attempts to assess the benefits and cost of conducting early intervention programs. The EIRI data were recoded at the item level for the Psychometrics Project, which established norms, reliability, and validity information on self-report, family-functioning measures gathered from families having children with disabilities.

The GT study results suggested that the items facet made a large contribution, indicating that there may not be any established trends in item responses. An explanation for the items facet indicates that the PSI forms provide an accurate measure of overall parental stress. According to the times facet results, the effects of
time are minimal except the increase between occasion one to occasion two. Classical reliability theory (CRT) and GT analyses provide contradictory results, probably due to GT’s multiple error source analyses compared to CRT’s examination of a single error source in one analysis.

GT study analyses indicate that the highest g and phi coefficients are produced with the highest number of administrations and items. However, administering the highest number of administrations and items would be impractical within any setting. The original number of items from the Parent Domain, Child Domain, and short PSI total score should be administered twice to increase the dependability of scores and still fall within practical limitations.

A researcher and/or practitioner may want information to decide what form, long or short, to choose. If the PSI is to be used as a quick screening tool or as one test in a complete assessment, the short form may be of more use. If the PSI is to be used as a primary source of information about parent and child interactive systems, the long PSI version would be recommended.
ACKNOWLEDGMENTS

I would like to express my appreciation to my chair, Susan L. Crowley, for guiding me through this laborious process and editing a countless number of revisions. In addition, Sue deserves my thanks for allowing me to utilize data from the Psychometrics Project. I would also like to thank Xitao Fan for helping me to understand the complexity of generalizability theory and answering my questions at my convenience. Mark Innocenti also deserves my thanks for keeping me on the right track in the early intervention literature. My committee as a whole unit was very supportive and patient with me in this process.

The Early Intervention Research Institute provided the data for this study, and the people responsible for the development and implementation of the EIRI studies (Contract #800-85-0173) should be acknowledged for their involvement in the EIRI project. Without EIRI data and access to those records, this project would not have been possible.

My mother, father, and brother deserve special recognition for being understanding and supportive of my goals and dreams. Without their care and support, I certainly would not be as successful as I am today.

My utmost appreciation goes to my wife, Donna, who moved to Utah so that I could attend graduate school. She left the comfort of family, friends, and a good teaching position so that I could continue my educational opportunities. She sacrificed time we could have spent together so that I could work on my thesis and other projects. In addition, her patience, tolerance, and wisdom assisted me during the many months of endless revisions.

Jim D. Sharpnack
# CONTENTS

<table>
<thead>
<tr>
<th>ABSTRACT</th>
<th>iii</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>v</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>viii</td>
</tr>
<tr>
<td><strong>CHAPTER</strong></td>
<td></td>
</tr>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>II. REVIEW OF LITERATURE</td>
<td>7</td>
</tr>
<tr>
<td>Parenting Stress</td>
<td>7</td>
</tr>
<tr>
<td>Parental Stress Instruments</td>
<td>10</td>
</tr>
<tr>
<td>Classical Reliability Theory</td>
<td>20</td>
</tr>
<tr>
<td>Fundamental Aspects of Generalizability Theory</td>
<td>25</td>
</tr>
<tr>
<td>Generalizability Theory Analyses</td>
<td>27</td>
</tr>
<tr>
<td>Applications of Generalizability Theory</td>
<td>33</td>
</tr>
<tr>
<td>III. METHODOLOGY</td>
<td>40</td>
</tr>
<tr>
<td>Sample</td>
<td>40</td>
</tr>
<tr>
<td>Procedures</td>
<td>42</td>
</tr>
<tr>
<td>Informed Consent Procedures</td>
<td>42</td>
</tr>
<tr>
<td>Measures</td>
<td>42</td>
</tr>
<tr>
<td>Analyses</td>
<td>46</td>
</tr>
<tr>
<td>IV. RESULTS</td>
<td>49</td>
</tr>
<tr>
<td>Parent Domain</td>
<td>49</td>
</tr>
<tr>
<td>Child Domain</td>
<td>54</td>
</tr>
<tr>
<td>Short Form</td>
<td>58</td>
</tr>
<tr>
<td>V. DISCUSSION</td>
<td>63</td>
</tr>
<tr>
<td>Review and Interpretation of the Results</td>
<td>63</td>
</tr>
<tr>
<td>Classical Reliability Theory and Generalizability Theory Analyses</td>
<td>71</td>
</tr>
<tr>
<td>Protocol Improvement</td>
<td>73</td>
</tr>
<tr>
<td>Limitations</td>
<td>75</td>
</tr>
<tr>
<td>Future Recommendations</td>
<td>77</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>79</td>
</tr>
</tbody>
</table>
APPENDICES

Appendix A: Institutional Review Board Approval .................................. 86
Appendix B: Parenting Stress Index--Long Form ................................. 89
Appendix C: Parenting Stress Index--Short Form ............................... 97
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Two Facet Design and Its Seven Variance Components</td>
<td>29</td>
</tr>
<tr>
<td>2.</td>
<td>Mother Demographic Characteristics</td>
<td>41</td>
</tr>
<tr>
<td>3.</td>
<td>Child Demographic Characteristics</td>
<td>41</td>
</tr>
<tr>
<td>4.</td>
<td>Facets and Percentage of Total Variance in the Parent Domain</td>
<td>50</td>
</tr>
<tr>
<td>5.</td>
<td>Parent Domain D Study Results for One, Two, and Three Administrations at Various Item Levels</td>
<td>51</td>
</tr>
<tr>
<td>6.</td>
<td>Facets and Percentage of Total Variance in the Child Domain</td>
<td>55</td>
</tr>
<tr>
<td>7.</td>
<td>Child Domain D Study Results for One, Two, and Three Administrations at Various Item Levels</td>
<td>56</td>
</tr>
<tr>
<td>8.</td>
<td>Facets and Percentage of Total Variance in the Short PSI</td>
<td>59</td>
</tr>
<tr>
<td>9.</td>
<td>Short PSI D Study Results for One, Two, and Three Administrations at Various Item Levels</td>
<td>61</td>
</tr>
</tbody>
</table>
CHAPTER I
INTRODUCTION

With the introduction of P.L. 99-457 and P.L. 101-476 (the Individual with Disabilities Act, IDEA), the Congress of the United States agreed to increase and improve services for young children with special needs and their families. According to early intervention research results, benefits from such programs included the enhancement of child development, increased capability of families to meet the needs of children with disabilities, reduction of institutionalization and educational costs, and improved proficiency at the state agency level to provide services for families having children with disabilities and their families (Ad Hoc Part H Work Group, 1995).

The Part H program, a major section of the IDEA legislation, focused on providing early intervention services to children with disabilities ages zero to three. The increased focus on young children with disabilities was a result of research suggesting that a child with special needs has more obstacles than a child without disabilities; these barriers may inhibit the child's ability to acquire basic skills (Ad Hoc Part H Work Group, 1995). Additionally, early intervention programs are based on the assumption that the probability for reducing or eliminating developmental problems is much higher during early childhood. Therefore, a primary goal for early intervention programs is the accurate identification of developmental problems and risks for delay when possible. To provide appropriate services to individuals who need them, proper identification of child health and developmental problems must be established (Ad Hoc Part H Work Group, 1995). The accurate identification of family resources and strengths, an integral part of providing services for a child with a disability, hinges on the proper assessment of family needs with reliable instruments.
Early intervention services assist family members in taking care of the needs of a child with a disability. Addressing the needs of a child with a disability, the traditional priority of service agencies, is an important component of evaluation. However, the examiner essentially only carries out half the investigation by not conducting further assessment of the family's needs.

Within the past few years, early intervention research and legislation have evolved from focusing solely on the needs of the child to examining the needs of the entire family system. The evaluation of the infrastructure of the family provides pertinent information about the utilized/unidentified supply of available resources. Assistance may be required from each family member to contribute to the caretaking needs for a child with a disability. For example, siblings may be asked to take on more responsibilities with younger siblings or household chores while parents attend to the needs of a child having a disability. Ultimately, the identification of family needs directly and indirectly addresses the needs of a child with a disability. Therefore, the extension of assessment to encompass the needs of the entire family serves to assist all family members and a child with a disability.

However, there are few instruments designed to assess families with children having disabilities; consequently, researchers and early intervention service providers often use measures that are most readily available. The use of existing research instruments poses a problem since most of these instruments were not validated on special populations. Research conventions demand that a research tool, such as a self-report assessment tool, must be validated on the pertinent characteristics of a specific population to be used in a research study (APA/AERA/NCME, 1985). For example, if a test had been standardized on a group of families having children with disabilities, other researchers could administer the test to families having children with disabilities.
However, if the same test had not been standardized or validated on families having children with disabilities, researchers should not administer the test to families having children with disabilities unless it is an attempt to provide validity or standardization information. Therefore, proper instrumentation is a fundamental requirement of competent research conventions. The practice of using evaluation instruments on populations who were not represented in the standardization process calls into question the dependability of test scores and the validity of the resulting research conclusions (Crowley, 1995).

There are many pertinent variables worthy of study in the examination of family needs, and parenting stress is an important area identified in the assessment of family needs. The assessment of parenting stress plays a key role in the evaluation of family needs and the needs of a child with a disability. Parental stress information gathered from assessments furnishes an introduction to the family's strengths. Hanson and Hanline (1990) stated that the identification of family strengths, such as personal coping resources and social supports, are pertinent for families having children with disabilities. Family resources can then be used to work on weak areas such as child behavior problems and/or dysfunctional parent behaviors.

The use of family-need instruments can provide a clearer picture of a family's resources and weaknesses. One of the most widely used tools to evaluate family needs in the research literature is the Parenting Stress Index (PSI; Abidin, 1990a). The PSI is a popular research tool because of its unique assessment of parent, child, and didactic stress and the strong psychometric background that serves as its foundation. Abidin's approach to the creation of the PSI emphasizes family needs, and the theoretical/research basis of the PSI was founded on mother-child systems under stress. In addition, Abidin (1990a) recognized the cumulative effects of stress on the mother/child
system and acknowledged not only multiple sources of stress but also the variation of the effects of persistent stress in contrast to short-term stress effects. For example, the long-term effects of stress associated with a child's illness would be more persistent than stress connected with an isolated incident such as a child receiving a temporary injury. Acknowledging the effects of long-term stress led Abidin (1990a) to create the PSI's factor structure around life events, the parent's sense of the child's activity level, and the parent's perception of being trapped by the parenting role.

A strong psychometric base is a key factor for any assessment tool like the PSI, and improvements within the field of psychometrics have produced increasingly powerful methods for use in test analysis. In traditional reliability analysis, classical reliability theory (CRT) evaluations examine similarity between scores of comparable content. Reliability analyses yield information describing various individual error components with correlation coefficients, and the effects of individual error components (e.g., time, test forms, and items) are examined in separate CRT analyses. Instruments with poor reliability scores produce undependable data. Assessing a "true score" is difficult when the evaluation is made using undependable data.

A recent improvement in the field of psychometrics has been the development of generalizability theory (GT). A major strength of GT includes assessment of several error sources in a single analysis. Despite the most advanced psychometric analyses, the effects of all variance components (e.g., random error and other sources of unsystematic variance) are unknown, and thus, no instrument is perfectly reliable across all situations. Although GT cannot perform the impossible task of identifying all error variance components, the inclusion of interactions and the assessment of multiple error sources in GT analyses makes this technique a more powerful means of assessing
score dependability than CRT analyses (Cronbach, Gleser, Nanda, & Rajaratnum, 1972).

Additionally, GT can be used in the design of protocol improvement. Information obtained from GT analyses provides evidence "to determine the best measurement design with which to get the most reliable scores in the most efficient manner" (Eason, 1991, p. 88). The results of GT can assist an examiner in improving the dependability of test scores by providing estimates of dependability with alterations in the number of testing occasions, test items, or forms. However, GT analyses can also assist researchers in weighing the consequences of gathering more test data, involving the expenditure of more funds, or stopping the data collection process because further assessment would provide minimal information given the study's costs.

As of yet, a study has not been conducted that has analyzed long or short form PSI data utilizing methodology based on GT. The application of GT to the PSI can yield critical information about known components such as the optimal number of administrations and items needed to increase score dependability. In addition, GT can be used to explore alternative measurement protocols using the PSI to obtain the most efficient and dependable data obtained from families having children with disabilities. With one exception (Crowley, 1995), there is a paucity of studies examining the reliability of PSI scores on families having children with disabilities. With the focus on early intervention and family functioning, the data from this study may be critical in making good developmental, financial, and policy decisions.

The purpose of this study was to provide detailed information about the dependability of PSI scores in families with children having disabilities using GT. This study is also designed to use GT to make information available regarding a guide to the process of maximizing methods used to obtain data. The foundation of this project rests
on three assumptions. First, appropriate evaluation must take place to meet and understand the essential needs of families having children with disabilities in research and in direct services. Second, the strong psychometric and research background of the PSI makes it one of the most widely used self-report instruments available to researchers. A self-report instrument with a strong psychometric and research background should be used in any study. Therefore, the PSI will be used in this study. Third, GT provides a more powerful method of assessing the dependability of test scores than CRT analyses. Therefore, the present study attempted to apply GT, the most efficient method of obtaining information on test-score dependability, to the PSI, one of the more psychometrically sound instruments that is currently available, to understand families better who confront disability issues on a daily basis. Specifically, the study will address the following research questions.

1. How dependable is the long version of the PSI Parent and Child Domain scores across 3 years using GT?

2. How dependable is the short version of the PSI across 3 years using GT?

3. What modifications can be made in the number of administered items and occasions in Parent Domain, Child Domain, and the short PSI form to increase dependability of test scores?
CHAPTER II
REVIEW OF LITERATURE

This review will summarize several domains of literature relevant to the present study. These topics include parenting stress, parental stress instruments, classical reliability theory, aspects of generalizability theory, and generalizability theory analyses.

Parenting Stress

The following section presents a review of the literature focusing on stress among families having children with disabilities. The review will cover the traditional views of stress in families having children with disabilities and current assumptions in the research literature. Past stress theories have typically highlighted pathological forces such as divorce and abuse in families having children with disabilities. The normality approach, a more recent stress theory, emphasizes family adaptations and coping skills.

Traditional Stress Research Views

Health and social findings are primarily responsible for the traditional, pathological assumptions of stress among families having children with disabilities. First, past research has indicated that families having children with disabilities have reported higher stress level than families who do not have children with disabilities (Hanson & Hanline, 1990; Kazak & Marvin, 1984). Second, the demands of caring for a child without a disability are not as great as meeting the needs of a child with a disability (Mahoney, O’Sullivan, & Robinson, 1992).

The pathological approach suggests that a child with a disability "produces" or "causes" high dysfunction within the family unit (Innocenti, Huh, & Boyce, 1992). Low social support (Kazak & Marvin, 1984), low socioeconomic status (Chetwynd, 1985),
high anxiety levels (Wyckoff & Erickson, 1987), high monetary expenditures (Holroyd, 1974), high rates of parental suicide and divorce (Price-Bonham & Addison, 1978), and significant vulnerabilities to family discord (Gallagher, Beckman, & Cross, 1983) were historically identified as some of the negative outcomes associated with high parenting stress in families having children with a disability. Past research asserted that children with disabilities appeared to be at greater risk for child abuse than children without disabilities (Gaines, Sandgrund, Green, & Power, 1978).

Despite the history of the pathological approach, problems exist within the studies that emphasize this particular point of view. Glidden (1993) reported that studies focusing on the pathological approach were plagued by characteristics such as lack of control groups or poorly selected comparison groups, unrepresentative sampling populations that are limited due to families who choose not to participate in the study, unknown psychometric characteristics of tests, lack of replication, and generalizations from one person's experience, typically, the mother, to the entire family's perception of life events.

The pathological approach continues to be a widely held view in research and interventions with families having children with disabilities. More specifically, some researchers have typically perceived families having children with disabilities as dysfunctional (Kazak, 1986). In addition, because of financial, emotional, and social obstacles, life stressors such as divorce, abuse, and financial stress exist within these families. Despite the fact that many families having children with disabilities do report higher stress levels than families having children without disabilities, the source of stress has yet to be identified due to studies that report conflicting evidence (Innocenti et al., 1992). In summary, the pathological approach represents a traditional view of
families having children with disabilities, and some of the studies exemplifying this point of view suffer from research flaws.

**Normality Method**

The normality method represents a departure from traditional research and applications in working with families having children with disabilities. This method grew out of the influences of the Individual with Disabilities Education Act (IDEA) and P. L. 99-457 (Innocenti et al., 1992). In the normality method, a more comprehensive approach in the perception of all families is identified. Stress is recognized to occur in all families, and the focus is on learning new coping skills and using resources within the family to lessen the effects of stress (Innocenti et al., 1992).

Abidin (1990a) defined parenting stress as child, parent, and life variables interacting with parental roles. Essentially, the term was defined so that it could be applied to all family systems in accordance with the normality assumption. However, in order to combat current stress levels, problem areas must be correctly identified with the family system. Assessing the level of parental stress among families having children with disabilities is crucial for the development of appropriate services to meet the needs of the child and family. By encouraging the reports of problem areas by parents, a beneficial solution to all family members can be the main intervention focus.

Recently, the normality method has offered a different view of parenting stress theories. Problems such as divorce, suicide, financial hardships, and so forth, which had been previously "caused" or associated with the disability as a result of the pathological approach, have been identified in all families. New forces within early intervention research have begun to focus more on coping skills and resources within families having children with disabilities. Attempts are being made to utilize already
existing resources and develop new skills that can ease the stress levels of families having children with disabilities. The focus of the normality method identifies stressors within all families. Families having children with disabilities have stressors due to life circumstances and not simply due to the fact that a child within the family has a disability.

Summary

Parenting stress is defined as the interaction of parent, child, and life variables (Abidin, 1990a). Past research dictated a pathological approach to describe families having children with disabilities by focusing on outcome variables such as suicide, divorce, and financial difficulties. By focusing on adaptive resources and skills, the normality method reflects an alternative to the pathological approach that has been recently more accepted within research literature. Abidin's (1990a) broad definition of parenting stress used in the development of the PSI reflects the normality method.

Parental Stress Instruments

Due to the importance of parental stress, accurate assessment is critical. In order to evaluate the effectiveness of early intervention programs, studies need to utilize the most accurate instruments with superior, psychometric properties. Due to the lack of information regarding psychometric characteristics, strengths, and weaknesses of many instruments, the following section will only focus on two instruments, Questionnaire on Resources and Stress (QRS) and the PSI. The PSI and the QRS have a thorough research history mainly due to the number of articles that have been published using these instruments. Although several other parental stress measures exist, thorough information on those measures was not available in the searches conducted by the
researcher. The review of stress instruments will address strengths and weaknesses, as well as descriptions of QRS and PSI long and short versions. More detailed descriptions for both PSI forms are contained in the Methods section.

Questionnaire on Resources and Stress

One of the primary parental stress inventories in the past 20 years was the Questionnaire on Resources and Stress (QRS; Holroyd, 1974). The QRS consists of 285 true-false items designed to measure 15 areas that relate to caregiving practices of families having children with disabilities. The three factors of QRS include parent, family-functioning, and child-related problems. The QRS has been used to differentiate between families having children with a variety of disabilities (Holroyd, 1974). For the 285-item scale, the Kuder-Richardson-20 coefficient is reported to be .96 (Touliatos, Perlmutter, & Straus, 1990). Touliatos et al. (1990) reported that the length of the QRS is an obvious limitation. Validity information for QRS scores was extrapolated from studies conducted on parents of children and adolescents (Touliatos et al., 1990).

Several studies have been conducted to test the discriminant validity of the QRS. Discriminant validity is used to assess the degree to which an instrument can separate or discriminate between two separate groups based on certain characteristics (Anastasi, 1988). A correlation coefficient between test scores is produced and serves as the determining factor. Studies with low coefficients are interpreted to show high differences between groups. However, high coefficients indicate that the administered test does not detect significant differences between groups. Holroyd (1988) provided extensive information on the discriminant validity of the QRS scores. Discriminant validity studies using the QRS include clinical versus nonclinical comparisons, clinical versus clinical comparisons, and child or parent variables to investigate the possible
moderating effects of those variables on stress. Examples of discriminant validity studies include the following investigations: differentiating between families such as autism versus Down syndrome versus psychiatric outpatient families (Holroyd & McArthur, 1976), intellectually disabled versus nondisabled families (Holroyd, Brown, Wikler, & Simmons, 1975), disabled versus heterogeneous disabled families (Friedrich & Friedrich, 1981), families with and without a child with neuromuscular disease (Holroyd & Guthrie, 1979), and between families having children who are intellectually disabled with and without cerebral palsy (Friedrich, 1979).

The cumbersome number of items in the QRS resulted in the development of three short forms of the QRS based on the results of factor analytic studies. However, the psychometric properties of the short QRS forms are not extensive (Scott, Sexton, Thompson, & Wood, 1989). One of the revised short forms of the QRS contains 52 items. The QRS-F (Friedrich, Greenberg, & Crnic, 1983) is a 52-item scale divided into four scales: (a) parent and family problems, (b) pessimism, (c) child characteristics, and (d) physical incapacitation, primarily used for the assessment of families having children with disabilities. Approximately 12% of the standardization sample for the QRS-F represented families without children with disabilities (Friedrich et al., 1983). The correlation between the QRS-F and the QRS was .99 (Touliatos et al., 1990). Internal consistency reliability, coefficient alpha, for the total score was .92 (Scott et al., 1989) and .95 (Friedrich et al., 1983). The construct validity of QRS-F was supported by the results of factor analytic studies (Scott et al., 1989).

The QRS-SF (Holroyd, 1982), a second QRS short form, contains 66 items and contains 11 subscales entitled dependency and management, cognitive impairment, limits on family opportunity, life span care, family disharmony, lack of personal reward, terminal illness stress, physical limitations, and personal burden for respondent. This
version of the QRS was developed to include factors pertaining to social support, financial problems, and social obtrusiveness lost in the QRS-F (1983) short version of the QRS (Salisbury, 1986). Kuder-Richardson reliability analyses for the subscales range from .34 to .84, and the overall internal consistency coefficient was equal to .79 (Salisbury, 1985).

The QRS-SFA (Salisbury, 1986) contains 48 items divided into seven scales with a various number of items in each scale. This version of the QRS was developed to include the factors left out in the QRS-F (1983) short version of the QRS and improve the psychometrics of the QRS-SF (1982). The seven subscales are entitled life span care, cognitive impairment, child characteristics, family disharmony, pessimism, physical limitations, and financial stress. Forty-one percent \( (n = 78) \) of the standardization sample included families having children without disabilities. Internal consistency coefficient for the QRS-FA was .76, and the coefficients for the subscales ranged from .59 to .84 (Salisbury, 1986). Salisbury (1989) provided information for the construct validity of the QRS-SFA by correlating QRS-SFA total score, scores on the seven subscales, and criterion measures including the Beck Depression Inventory (BDI; Beck & Beamesdorfer, 1974), Lock-Wallace Marital Adjustment Inventory (Locke & Wallace, 1959), and Family Social Support Scale (Dunst, Jenkins, & Trivette, 1984), physical needs, income, and level of functioning. The QRS-SFA Total score significantly correlated with the BDI (.43) and physical needs (.22).

The Clarke modification of the QRS is a 78-item test designed to be used specifically with families having severe childhood psychopathology such as autism and developmental disabilities such as intellectual disabilities (Konstantareas, Homatidis, & Plowright, 1992). Nine subscales for the Clarke modification of the QRS include child characteristics, community reaction, time demands, family sharing, presenting
symptoms, sacrifice/martyrdom, supports, family enrichment, and existential issues.

Split-half reliability coefficients for this QRS version produced .85 for the two halves of the test, and .89 for items with odd and even numbers. Construct validity analyses examined the relationship between semistructured reports of global stress against total stress scores obtained from the Clarke QRS. Lower stress reports correlated more highly with Clarke QRS scores than moderate or high stress reports (Konstantareas et al., 1992). Discriminant validity analyses consisted of contrasting the scores between mothers and fathers of autistic, learning disabled, intellectually disabled, and a control group of children (Konstantareas et al., 1992). Discriminant functions distinguished among the groups of children and mothers versus fathers responses.

In addition to the short forms already mentioned, Glidden (1993) indicates that two other short versions of the QRS are being formulated. Scott et al. (1989) developed a version of the QRS that has a slightly different factor structure than the original QRS. Glidden also stated that Engelhardt (1990) is developing a form of the QRS that would be specifically used for the guardians of infants.

Although the QRS and its short versions represent some of the most widely used instruments employed to measure stress in families having children with disabilities (Wikler, 1986), the instrument has several problem areas. Clayton, Glidden, and Kiphart (1994) pointed out that the QRS items are a mixture of child characteristics and family functioning level. However, the scores from the child and family domains are not separated but are added into a single total score. Therefore, a parent having a child with a disability can produce a high score solely due to the physical and cognitive limitations of the child. Consequently, a QRS profile may manufacture a high score that may be presumed to exist because of family dysfunction, but in actuality, the family may have an adequate functioning level. As Glidden (1993) pointed out, the existence of
limitations due to a disability does not indicate the presence of maladjustment or pathology. Therefore, interpretations of the QRS forms about family dysfunction should be made with great caution.

The 285-item QRS is not widely used because of its long administration time (Scott et al., 1989). Friedrich et al. (1983) also pointed out that the original QRS has weak psychometric properties with low internal consistency reliability and validity as well as the lack of factor analytic studies using the QRS. The QRS was standardized in three pilot studies on 64 parents having children with disabilities (Holroyd, 1974). The QRS-FA had a larger and more representative standardization sample than the original QRS (Salisbury, 1986). An instrument with such a small sample may not yield a representative sample of families having children with disabilities.

Another problem with the QRS is deciding which of the seven forms to choose. Therefore, the researcher must pick between the original QRS and the six QRS short forms, which can be a daunting task. Although the questions from the original QRS are used in the short forms, each of the short forms specifically address different areas. A comparison of the QRS-SF, QRS-F, and the QRS-FA produced incongruent results for family functioning levels despite the similarities among the instruments (Glidden, 1993).

Parenting Stress Index

Since its introduction in 1983, the PSI has become a widely used tool in family research. Abidin (1990a) developed the PSI to examine parental stress as defined by child characteristics, parent variables, and life stresses associated with the parental role. The PSI has been most utilized for early identification screening, individual diagnostic assessment, pre-post measurement of intervention effectiveness, and evaluating effects of parental stress on other parental and behavioral variables. Unlike
the QRS, the PSI was not standardized on families having children with disabilities. A long and a short form were developed for the PSI. In this section, psychometric information for both PSI forms will be addressed, and both forms will be thoroughly discussed in the Methods section. A breakdown of the PSI subscales for both forms will also be provided in the Methods section.

**PSI (Long Form)**

The PSI (Abidin, 1990a) is a 101-item self-report instrument that evaluates parenting stress associated with parent and child variables. One asset of the PSI is its strong psychometric integrity, and information on reliability and validity for the PSI can be found in the PSI manual. Abidin (1990a) reported that internal consistency coefficients range from .62 to .70 for the Child Domain subscales, from .55 to .80 for the subscales of the Parent Domain, .89 for the overall Child Domain scale, and .93 for the overall Parent Domain scale. Zakerski (as cited in Abidin, 1990a) computed the following test-retest reliability information across a 3-month interval for 54 parents: Child Domain, .55; Parent Domain, .69; Total Stress Score, .88. Test-retest coefficients for the Child Domain (.55), Parent Domain (.70), and the Total Stress Score (.65) remained fairly stable in a study by Hamilton (as cited in Abidin, 1990a) for 37 mothers across a 1-year period. Overall, the internal consistency and test-retest reliability information are similar or better than those reported for the QRS.

The PSI manual provides thorough details of literature supporting the validity of PSI scores. Concurrent validity studies were conducted by computing a correlation coefficient between two test scores that claim to measure the similar constructs (Anastasi, 1988). Information exists on studies examining concurrent validity between PSI scores and the Bayley Infant Development Scales (Bayley, 1969), Child Behavior
Checklist (Achenbach, 1991), and the State-Trait Anxiety Index (Spielberger, 1983). A study by Zakreski (as cited in Abidin, 1990a) found that the Bayley Infant Development Scales significantly correlated with the PSI total score at 3 ($r = .42$) and 6 months ($r = .66$) postpartum for parents of full- and preterm babies. In an examination of parents of preschool children with and without histories of serous otitis media, Casey (as cited in Abidin, 1990a) found statistically significant relationships between the Achenbach total score and the PSI Parent ($r = .56$) and Child Domains ($r = .40$). Oppenheimer (as cited in Abidin, 1990a) found that the Child Behavior Checklist Total Score and the PSI Child Domain score had a statistically significant relationship ($r = .80$). Jenkins (as cited in Abidin, 1990a) discovered statistically significant relationships between the PSI total score and the State Anxiety ($r = .71$) and Trait Anxiety ($r = .84$) of the State-Trait Anxiety Index. Overall, there is strong supporting evidence for the concurrent validity of the PSI.

Discriminant validity studies reported in the PSI manual include many investigations. A study by Chavkin (as cited in Abidin, 1990a) found that the Child Domain discriminated statistically significantly between families having children with spina bifida, no disability, and autism. A study by Jenkins (as cited in Abidin, 1990a) found statistically significant differences between parents of intellectually handicapped, emotionally disturbed, and learning disabled children using Child Domain scores. Tam, Chan, and Wong (1994) used the PSI to successfully discriminate between Hong Kong mothers with high stress versus low stress with an accuracy of 93%.

Predictive validity is a method used to predict the likelihood of an event or the development of a diagnosis over time (Anastasi, 1988). Accurately predicted events or diagnoses result in high predictive validity, and false positives result in low predictive validity. In a study by McGaughey (as cited in Abidin, 1990a), the PSI was
administered to 85 parents when the children were 8 months old, and the Child Behavior Problem Checklist was administered to teachers when the children reached age 5. For the parents whose PSI total score reached the cutoff rate, which Abidin (1990a) described as a situation requiring professional intervention, the overall hit rate using CBPC criteria for anxiety and behavioral problems with the students was 91%, the false positive rate was 29%, and the false negative rate was 5%. In a study of children identified with developmental delays in an early intervention program, Upshur (as cited in Abidin, 1990a) reported that using the PSI's 90th percentile as a cutoff rate accurately identified 89% of the parental/student relationships requiring professional intervention.

**PSI (Short Form)**

Some researchers prefer shorter instruments to reduce administration time for clients and reduce scoring time for practitioners. Therefore, Abidin developed the PSI/Short Form (PSI/SF) (1990b) to assess parent-child interactions in a short time period. While the long version of the PSI provides more information on parent-child variables, the shortened form requires less time while still providing pertinent data on the parent-child relationship.

Although research on the PSI/SF is not as extensive as the PSI's longer version, the research that does exist on the short form continues to support the PSI's strong psychometric integrity. Test-retest reliability information over a 6-month period revealed scores of .84, .85, .68, and .78 for total, parental distress, parent-child interaction, and difficult child factors, respectively. Internal consistency coefficients continue the high reliability trend with scores of .91, .87, .80, and .85 for the total, parental distress, parent-child interaction, and difficult child scores, respectively.
Summary

The PSI and QRS represent prominent self-report inventories of parental stress in the research literature. Although other instruments exist to measure parental stress, there is a lack of information regarding their psychometric characteristics, a fundamental point discussed in this study, making them less useful. The QRS (Holroyd, 1974) was developed to measure stress in families having children with disabilities. The length of the original QRS, small standardization sample, and its psychometric characteristics led to the development of several shorter forms of the QRS. The existence of six short QRS versions leaves the researcher the perplexing question of choosing among them. Although the short versions contain original questions from the QRS, a comparison of three of the more popular short QRS versions produced conflicting results (Glidden, 1993).

The PSI is one of the most widely used instruments within the research literature. Because of its dominance in the research literature, a large amount of research has focused on its psychometric characteristics and applications. The psychometric integrity of the PSI scores is well documented due the extensive classical reliability and validity studies. The PSI is also widely accepted in practical settings where it is used as a diagnostic, screening, and research instrument among a variety of sources such as early intervention agencies and pediatric clinics. The standardization sample for the PSI did not represent families having children with disabilities. However, attempts have been made to validate PSI forms on families having children with disabilities (Crowley, 1995; Innocenti et al., 1992; Smith & Innocenti, 1993), and the PSI can be used with this population. The PSI stands above the other instruments because of its widespread use in the literature and applicable settings, strong psychometric background, and relatively quick administration.
Classical Reliability Theory

The following sections contain information about definitions of common terms and variance components associated with classical reliability theory (CRT), a description of the different types of reliability analyses based on classical test theory, and the strengths and weaknesses associated with CRT.

Definitions of Terms Used in Classical Reliability Theory

Reliability. Reliability refers to the degree of consistency or dependability of test data. In order to assess dependability, a researcher must examine variance within a data set. Variance can be simply defined as the "spread" or distribution of test scores in a sample. In CRT, three forms of variance are considered: observed, true, and error components (Traub, 1994). Scores that tend to be "close" in number (i.e., 50 scores from 30-60 points on a 100-point scale) tend to have a smaller variance as opposed to scores that have a large separation (i.e., 50 scores from 0-100 on a 100-point scale). Different types of reliability will be discussed later in this section.

True score variance. True score variance represents subject variation due to treatment or actual differences. An example of true score variance is exemplified by test score differences for subjects in two groups, one that has had a precatory study course before an exam and another who did not take the course. If the study course group does better on the exam in comparison to the other group, part of the explanation may be attributable to the effects of the study course in addition to other variables.

Error variance. Error variance comprises known or unknown factors that extend beyond the researcher’s control, such as illness or fatigue. Eason (1991) stated that error variance exists, but the extent of its effect is obscure at best, since all factors are
not known. Error variance is multifaceted because it represents all variables that affect outcomes, with the exception of actual differences. To continue the previous example, suppose an unprepared student scored better than the mean average of the students in a precatory course. The unprepared student may have gotten a better score due to getting plenty of rest the night before the exam, exhibits better time management skills than other students, or has excellent test-taking skills. The previously mentioned reasons for the student's success could be attributed to error variance. However, a variable that is considered to be an error component in one study may not be considered an error component in another study. For example, if a researcher is examining the effects of rest on test performance, sleep activity the night before the exam may be considered as an independent variable.

**Observed variance.** The fundamental principle of CRT states that the addition of error and true score variance is equal to observed variance, and observed variance is ultimately what the researcher measures (Traub, 1994). Observed variance will never equal true variance because of the influences of the unknown or known error components. In the testing example, the effect of the course and factors such as motivation, fatigue, or personal interest combine to make up the observed variance component.

**Classical Reliability Theory Analyses**

Although error variance is multifaceted, reliability analyses in CRT always measures a single source of error at a time. Test-retest reliability assesses observed variance associated with time, internal consistency analyses measure observed variance associated with item variance, and parallel forms estimate observed variance related to test forms. Each of the analyses will be discussed in the following section. In
CRT, a correlation coefficient is derived between two different test scores, indicating the degree of stability between the scores.

**Split-half reliability analyses.** In the split-half reliability procedure, a test is divided into two equal halves, and a correlation coefficient is computed between test scores from both halves. Split-half reliability coefficients are used to determine the consistency of the items from both test halves. Since the split-half coefficient produces low estimates of the reliability of the entire test, the Spearman Brown prophecy formula produces a coefficient reflective of the full-length test (Crocker & Algina, 1986).

**Internal consistency analyses.** Internal consistency procedures examine item homogeneity used to measure the content domain of items (Crocker & Algina, 1986). High-item homogeneity implies that all items of a test measure a similar content area. The internal consistency coefficient is computed by the average of all possible split-half coefficients. Kuder-Richardson reliability and coefficient alpha are the two available methods used to determine internal consistency procedures. The sole difference separating the two methods is determined by the number of available answers. More specifically, dichotomous answers (true-false, right-wrong) are only used to determine the Kuder-Richardson reliability coefficient, and coefficient alpha can use multiple response methods such as those used in Likert scales.

**Test-retest analyses.** To test for the effects of time, CRT computes a correlation between the score from Time 1 and the score from Time 2 on an instrument. A subject who takes the same test twice with an interval of time between the tests is likely to produce different scores due to various situational factors (e.g., subject’s health, climate, noise). A high test-retest reliability coefficient indicates similar results from Time 1 to Time 2, and test scores would appear to remain constant despite the effects of time. Test scores with high test-retest reliability appear more stable over time as
opposed to unstable test scores with poor test-retest reliability. Test-retest reliability is
dependent not only on the test structure but also on the construct measured by the test.
Poor test-retest reliability can be a result of poor structure and/or an unstable construct
measured by the test. The researcher is interested in the examination of the
consistency of responses over time.

Alternate form reliability analyses. Subjects who take the same test again may
recall previous responses or benefit from practice. To combat those effects, two test
forms measuring the same content can be developed. Alternate form reliability
examines time effects (test-retest) and item content (internal consistency) on test
scores. A single correlation coefficient between the test forms is computed to measure
the combination of two error forms in alternate form reliability. A high alternate form
reliability coefficient would indicate high consistency between alternate test forms with
similar content.

Strengths. The strengths of CRT include basic language and principles because of
its popular use. CRT refers to consistency in all fields, and all scientific fields utilize
CRT in research studies. CRT terms are straightforward and not difficult to
comprehend. Another strength of reliability is its relatively simple computation formulas
used to derive reliability coefficients.

Limitations. Estimates of error, and thus reliability information, vary in accordance
with the design of the analysis in CRT (Shavelson, Webb, & Rowley, 1989). A single
CRT analysis cannot simultaneously measure the contribution of time, item
homogeneity, and parallel forms. CRT analyses are criticized for only reflecting
variation in scores at a specific point in time (Traub, 1994). Testing a subject on more
than one occasion with the same test (test-retest reliability), with a parallel form on a
different occasion (alternate form reliability), or with items on the same test (split-half
reliability) provides not only different concepts of consistency but also independent numerical estimates (Traub, 1994). Therefore, reliability coefficients from CRT analyses can produce contradictory results (Eason, 1991). For example, scores may have high test-retest coefficients, low internal consistency coefficients, and low alternate-form coefficients.

One unwritten assumption in CRT analyses is that all variables are mutually exclusive. Mutually exclusive variables are classified as variables that do not overlap, or variables that can be classified in only one group (Toothaker, 1986). Therefore, CRT analyses ignore variable interactions and focus simply on mutually exclusive variance. The combination of error variables does not simply equal the amount to which two variables overlap, but extends beyond the merger of the combined variables (Toothaker, 1986). This new combination of variables forms another potential error source, interaction, which is not considered in CRT analyses.

Summary

The main focus of CRT is the dependability of test scores. Error, true, and observed variance are fundamental components of CRT. CRT analyses consist of test-retest reliability, alternate form reliability, or split-half reliability. The major strength of CRT consists of its wide range of applications across various research fields and its basic principles. However, CRT limitations include separate analyses that may produce contradictory results and the view that variance is not cumulative. Because of an assumption that variables do not overlap, CRT analyses ignore the possibility of interaction and cumulative effects between variance forms.
Fundamental Aspects of Generalizability Theory

Generalizability theory (GT) provides researchers with methods that can be used to overcome weaknesses associated with CRT. Like most theories, GT consists of a set of complex theoretical assumptions and a new language with unfamiliar terms. Therefore, this next section presents common definitions associated with GT, a conceptual overview of GT, a comparison of GT and CRT, and limitations of GT.

Generalizability Theory Definitions

Facet. Procedures used in GT analysis utilize two key terms, facets and a universe. A facet is simply a source of error variance, or according to Brennan (1983), a facet is a collection of comparable measurement conditions. Examples of commonly examined facets include items, tests, occasions, raters, or observers (Eason, 1991). The scenario of a researcher interested in examining the effects of occasions (i.e., time) and items on test scores is an example of a two-facet designed study. GT allows for interaction between various components. Therefore, in the previous example, occasions x items, occasions x persons, items x persons, and occasions x items x persons + error represent possible interactions among variance components.

Universe. Brennan (1983) described a universe as conditions for measurement or the situation under investigation. Shavelson and Webb (1991b) made a distinction between universe of admissible observations and the universe of generalization. The distinction between the terms relies on the breadth of generalization that a researcher wishes to use. For example, if a researcher wanted to generalize across occasions, the universe of admissible observations would encompass a broad perspective examining the dependability of all tests across all situations within a specific content area such as chemistry or psychology. However, the universe of generalization is more suited to the
specifics of the situation. If we continue to use our two-facet example, the universe of generalization may encompass the dependability of test scores of Psychology 101 students across years. The universe of admissible observations could include the dependability of any acceptable items taken at several points in time.

The group the researcher is testing refers to the objects of measurement. The objects of measurement differentiate subjects based on the facet of interest. Objects of measurement are the "entities about which the researcher wishes to make generalizable statements" (Eason, 1991, p. 89). The objects of measurement typically refer to persons and are not considered a facet of the design (Brennan, 1983).

**Generalizability Theory Benefits**

GT builds on CRT by recognizing and estimating the magnitude of the multiple sources of measurement error (Shavelson et al., 1989). CRT only focuses on one source of error at a time thereby ignoring other possible error sources. GT examines several possible error sources in a single analysis. Reducing the amount of error in data collection to assist in protocol improvement and identifying number of items needed to increase dependability are examples of multiple interpretations that are available to the researcher as a result of GT analyses (Shavelson & Webb, 1991a, 1991b). GT broadens CRT's research question from "What is the accuracy of the relationship between true and observed scores?" to "How precisely can an observed score of behavior generalize across situations?" In addition, GT analyses provide components that examine the interaction of the variables in a single study.

**Summary**

Facets refer to a source of variance in GT studies, and commonly used facets include occasions (test-retest reliability), items (split-half reliability), forms (alternate
form reliability), and the interaction of those components. A universe provides a researcher the opportunity to define the application of GT results to a wide range of events, the universe of admissible observations, and the specifics of the research investigation, the universe of generalization. Objects of measurement is a term that refers to the subjects who took the tests and therefore produce the actual test scores used in analyses. Benefits of GT include the examination of interaction facets and optimizing dependability with the increase and/or decrease in the levels of facets.

Generalizability Theory Analyses

The next section provides the reader with a more indepth examination of GT analyses. Variance components are outlined with examples presented for greater clarity. The two types of GT studies, G and D investigations, are introduced to familiarize the reader with both study types.

Variance Components in Generalizability Theory

If the researcher simply had to examine the effects of one variable such as testing occasions, CRT analyses would be suitable. However, more than one source of variability must be considered since variables are not mutually exclusive. GT assumes that variables are not mutually exclusive and error variance is assumed to overlap. In a G study, Shavelson and Webb (1991a, 1991b) stated that the number of variance components is determined by the number of facets in a study. For example, three sources of variability exist in a one-facet study of occasions: occasions, persons, and persons x occasions + error.

Consider the following scenario. A researcher examining occasion dependability (i.e., test-retest reliability in CRT) is using college students as the objects of
measurement. An interval between tests, such as pretest/posttest design, is used to address the effects of time and thus constitutes the variance component of occasions. The variance component of items is represented by all of the test items. The final piece, randomness, accounts for the last unexplained source of variance. With the exception of objects of measurement, all other variance components are used to examine the error facet (Shavelson & Webb, 1991a, 1991b). The researcher must contend with uncontrollable components such as random sources of variance. Therefore, random factors along with the interaction between scores and occasions x persons + error are joined into one category defined as error (Shavelson & Webb, 1991a, 1991b). The error category represents the variance that cannot be individually explained by scores and time.

The number of variance components is dependent on the number of facets in a study. If, as in the previous example, the researcher was interested in not only the effects of occasions but also items, the result would be a two-facet study. The objects of measurement would continue to be persons, and the facets would include occasions and test items. The interaction components would consist of persons x occasions, persons x items, and occasions x items. The error category would include the interaction of scores x time x items + personal history + randomness. Table 1 displays a two-facet design and its seven variance components.

Generalizability Studies

In GT, two study types, Generalizability (G) and Decision (D), are available to the researcher. G study results attempt to explain as much information about measurement error sources as possible. The first priority of a G study is to define the facets in order to define all possible sources of error. The purpose of a G study is to identify all
Table 1

Two Facet Design and Its Seven Variance Components

<table>
<thead>
<tr>
<th>Facets</th>
<th>Components</th>
</tr>
</thead>
<tbody>
<tr>
<td>occasions</td>
<td>(o)</td>
</tr>
<tr>
<td>occasions x items</td>
<td>(o x I)</td>
</tr>
<tr>
<td>persons x occasions</td>
<td>(p x o)</td>
</tr>
<tr>
<td>persons</td>
<td>(p)</td>
</tr>
<tr>
<td>persons x items</td>
<td>(p x I)</td>
</tr>
<tr>
<td>items</td>
<td>(I)</td>
</tr>
<tr>
<td>persons x occasions x items + error</td>
<td></td>
</tr>
</tbody>
</table>

An estimated variance component is produced for each facet in a G study and describes the estimated magnitude each facet donates to measurement error.
(Shavelson et al., 1989). The estimated variance components are added together to calculate the total variance component. The sum of all variance components equals 1.00, and the sum of all variance component percentages equals 100%. A percentage of the total variance is then computed for each estimated variance component. The computed percentage represents the amount each estimated variance component contributes to the total variance component. An example can provide a clearer understanding of this process. For example, the estimated variance component for the item facet is equal to .3, the occasions facet is equal to .2, and the objects-of-measurement facet is equal to .1. The items x occasions interaction is equal to .1, the occasions x objects interaction is equal to .1, the items x objects interaction is equal to .1, and the error facet is equal to .1. The following represent the amount of variance overlap between the total and each facet: item facet (30%), occasions (20%), objects (10%), items x occasions (10%), occasions x objects (10%), items x objects (10%), and error (10%). The variance component that accounts for the most variance is the items variance component.

**Decision Studies**

Whereas the G study defines measurement procedure, the D study takes that information (e.g., the estimated variance components) and attempts to modify measurement procedure to improve the dependability of each facet (Shavelson & Webb, 1991a, 1991b). In order to conduct a D study, the researcher must define the universe of generalization, interpret the data, and then apply the results for more efficient measurement procedures (Shavelson & Webb, 1991a, 1991b).

The process of a D study continues along the following series of steps. The first step in a D study is to define the universe of generalization, which includes the
identification of error facets such as occasions, forms, and so forth. Although a D study may or may not include all facets used in the G study, the D study cannot add facets that were not included in the G study (Eason, 1991). The measurement design of the G study is altered in a D study to vary facet conditions to investigate protocol improvement (Eason, 1991). Several D studies can be investigated on the sole basis of one study. Researchers are allowed to vary facet conditions to optimize the dependability of obtaining data. For example, an instrument has a total of 40 items, and a researcher wants to find the best conditions in which to administer the instrument by manipulating the number of items and occasions. A D study may provide information that leads the researcher to believe that the most dependable set of data can be obtained by administering 20 items twice versus 30 items in one administration or 10 items in three administrations. Any combination of facets can be used to investigate the impact on dependability of test scores.

Next, the researcher must choose how to interpret the data. The two choices for data interpretation include making relative decisions or absolute decisions. Relative decision interpretations focus on the rank comparison of individuals and result in a G coefficient. “Phil scored better than two thirds of his class on the SAT” would be an example of a relative decision. Absolute decisions highlight the individual’s performance, and phi coefficients are considered absolute decisions. An example of an absolute decision encompasses state teaching certificate guidelines that mandate that students correctly answer a certain percentage of questions in order to be certified. CRT analyses can only perform relative decisions and cannot be utilized in making absolute decisions. Like CRT studies, coefficients are produced in GT studies. The G and phi coefficients in GT exemplify how accurately an observed score can be applied across a set of similar situations. Phi coefficients cover a wider range of variability than
G coefficients and are thus consequently lower than G coefficients. G coefficients include the interaction of variance components, and phi coefficients include the interaction and other factors of variance components. Phi coefficients can be equal to G coefficients, but phi coefficients can never be greater than G coefficients. CRT and GT coefficients share some similarities in that both coefficients attempt to identify the amount of error that is due to true variance, and both tools have different applications, which ultimately depend upon the examination of error (Shavelson & Webb, 1991a, 1991b). Both the generalizability and phi coefficients are interpreted like reliability coefficients. The coefficients range from .0 to 1.00, and desirable coefficients fall between .80 and 1.00.

Finally, the researcher applies the results to make alternative designs in future research or protocol improvement to minimize error and maximize reliability (Shavelson & Webb, 1991a, 1991b). In a D study, estimated variance components from the G study are used to calculate alternative D studies. Each of the D studies includes estimated generalizability coefficients and phi coefficients. For all available items and occasions, a D study can compute the generalizability and phi coefficients for all available pairings between the various levels of occasions and items. The results of a G study with occasions and items as facets could be used in a D study to investigate methods to improve the dependability of all occasions and item facets. For example, the results of a D study may indicate that using only 50 items on two testing occasions may produce the more dependable testing scores than 100 items in one administration. Typically, as items and occasions increase, the generalizability and phi coefficients increase. Protocol improvement is the cost-effective component of GT in that it seeks to maximize the utility of data derived from the instrument. The researcher must weigh
when the increased dependability of scores no longer offsets the increases in time and money for participants, researcher, and agencies.

Summary

The interaction of variables prompts the use of analyses that assess variable overlap. GT analyses assess the contribution of each facet as well as the interaction of those facets. Variance components represent the amount of shared variance between each facet and the total amount of variance. G and D studies represent the types of investigation in GT. In a G study, facets and the universe of admissible observations are defined. Estimated variance components are computed for each facet in a G study. In a D study, the universe of generalization is defined, decisions are made on how to interpret the data, and procedures to improve measurement dependability are discussed. Relative decisions represent the rank comparison of individuals, and absolute decisions solely reflect individual performance. G coefficients are products of relative decisions, and phi coefficients result from absolute decisions. These coefficients represent the dependability of measurement, which is similar to the interpretation of CRT coefficients. Several D studies can be conducted in order to optimize measurement conditions. The researcher must weigh the benefits of measurement dependability and the cost of administration.

Applications of Generalizability Theory

A search of PSYCHLIT, Dissertation Abstracts, and ERIC databases from 1974 to 1996 resulted in over 100 studies utilizing GT. However, many of these GT studies evaluate interest areas not pertinent to the current study such as interrater reliability. In order to be included in this review, GT studies had to meet the three criteria. First,
studies must include evaluations of published instruments. Many GT studies utilize rating scales that were developed specifically for the study. These impromptu rating scales do not have a wide psychometric history or even a test manual that describes the scale characteristics. Given the focus on the PSI in the present study, GT studies that include published instruments will have higher comparative value. Second, reviewed studies must contain G and D study information. Some studies only contain G studies and lack D study information. Since the current study will include D study information, studies with G and D analyses will be discussed. Third, for comparison purposes, studies that provide information on variance components will be discussed. If CRT information is available, it will also be mentioned. However, information on two other GT studies that do not meet the criteria will be mentioned briefly.

The studies discussed in this section represent research similar to the present study. This section will focus on four studies that utilize assessment instruments with psychometric history and contain G study information with variance components and D study information. Some studies also conducted CRT analyses.

Studies with Fundamental Comparison Characteristics

The first study was conducted by Thompson and Melancon (1987) using the Group Embedded Figures Test (GEFT; Witkin, Oltman, Raskin, & Karp, 1971) with 175 undergraduate students. The subjects were randomly selected from students in mathematic courses, and each subject completed the GEFT in compliance with test manual instructions (Thompson & Melancon, 1987). The facets included two GEFT test sections, nine items each nested within both sections of the GEFT, and the interaction terms. Results from the G study indicated that the two highest variance components included the interaction of persons x items (nested within sections) and persons.
Sections of the test and the interaction of sections and persons were the lowest variance components. D study variations revealed that the administration of the GEFT in its present format produced one of the highest generalizability coefficients (.88) in comparison to several alternative formats. The addition of nine items produced a G coefficient equal to .92, and the addition of another section with nine items produced a G coefficient equal to .91.

Crowley, Thompson, and Worchel (1994) applied GT to the scores of 164 children on the Children’s Depression Inventory (CDI; Kovacs, 1986). Facets of the study included two testing occasions, the 27 CDI items, and the interaction terms. The persons x items x occasions + error accounted for 58% of the total variance, persons x items accounted for 17% of the total variance, and the persons component accounted for 13% of the total variance. Items accounted for 5.8% of the total variance, persons x occasions accounted for 5.2% of the total variance, occasions accounted for .2% of the total variance, and occasions x items accounted for .1% of the total variance. CRT analyses revealed that internal consistency coefficients for Time 1 were equal to .86 and .88 for Time 2. Test-retest reliability information produced a coefficient equal to .66. With 27 items, testing across occasions yielded G coefficients equal to .63 for one occasion, .75 for two occasions, and .81 for three occasions. With 54 items, testing across occasions yielded G coefficients equal to .67 for one occasion, .79 for two occasions, and .84 for three occasions. Information from CRT and GT produced different information possibly due to the interaction of variance components in GT. The increase in G and phi coefficients as a result of the D studies did not provide appreciable improvement for the alternation of the CDI’s present 27-item format. However, Crowley et al. reported that more reliable data can be collected from multiple assessments of depression scores over multiple occasions.
The System to Plan Early Childhood Services (SPECS; Bagnato & Neisworth, 1990) instrument is used to design and evaluate early intervention services for children between the ages of 2 and 6 who are developmentally delayed. Suen, Lu, Neisworth, and Bagnato (1993) investigated the ratings of 467 children by 467 parents and 30 professional teams. Two hundred sixty-two children had been identified as developmentally delayed, and 205 children served as control peers. Some professionals only rated some of the children, and all of the parents rated all of the children. Therefore, raters were nested within subjects (i.e., different children are rated by different parents and professionals). The number of raters in the D study ranged from one to four. The total number of items used in each of the D studies was 19. In the G study, the two highest predictors of total score variance were persons (18%) and the interaction of persons x raters x items + error (25%). With an increase in raters from one to four, D study analyses revealed that G coefficients ranged from .58 to .83, and phi coefficients ranged from .57 to .81. Although SPECS is typically administered by one rater, the conclusion by the researchers indicates that the most dependable set of SPECS scores is derived by using four raters. In most scenarios, Suen et al. concluded that SPECS scores had average reliability levels.

The Generalized Contentment Scale (GCS; Hudson, 1982) is a 25-item questionnaire designed to measure depressive symptoms. The Index of Self-Esteem (ISE; Fischer, 1978) is an evaluation of problems resulting from low self-esteem. Nugent (1994) investigated the dependability of the clinical cutoff scores on the GCS and the ISE. Three hundred forty university students and outpatients receiving services at a community mental health center completed the 25-item GCS scale, and 127 university students, community members, and outpatients completed the 25-item ISE. Separate analyses were conducted for both of the instruments, and both studies
consisted of a one-facet design. The variance components for both studies included persons, items, and persons \( \times \) items + error. For both studies, persons, and persons \( \times \) items + error contributed the most to the total variance component. For the GCS, the persons component accounted for 37% of the total variance, and the persons \( \times \) items + error term accounted for 55% of the total variance. For the ISE, the persons component explained 38% of the total variance, and the persons \( \times \) items + error explained 53% of the total variance. For the 25 items and 340 persons in the GCS analyses, the G coefficient was equal to .94, and the phi coefficient was equal to .93. For 25 items and 126 persons, the G coefficient was equal to .94, and the phi coefficient was equal to .93. An increase from 25 items to 30 items in both instruments produced a slight improvement of phi coefficients to .93 for the GCS and .93 for the ISE. The authors concluded that the GCS and ISE are clinical instruments that provide dependable test data resulting in the identification of clinically significant problems.

**Studies Without Fundamental Comparison Characteristics**

Many GT studies focus on different aspects from the present study, such as interrater consistency, or fail to report information that is pertinent this study, such as variance components, G coefficients, or D study analyses. The following set of GT studies met some but not the full three criteria and will therefore only be briefly discussed.

Conger, Conger, Wallander, Ward, and Dygdon (1983) applied GT to the seven-item hyperactivity subscale of the Conners' Teacher Rating Scale - Revised (CTRS-R; Goyette, Conners, & Ulrich, 1978) for the observations of 66 children by teachers and classroom observers in 13 classrooms at three schools. For the combined analyses, teacher and observer scores were combined into a component entitled mode. For the
combined analyses, the four largest variance components were school x mode (24.9%), mode (23.4%), persons x mode (21.2%), and persons x mode x occasion (17.2%). The school x mode effect was due to large variability between teacher and observers in different schools. The mode effect indicated that teacher scores had lower variability than observer scores. The persons x mode effect indicates that children ratings were variable according to teacher and observer scores. The persons x mode x occasions effect indicated that children ratings were variable across raters and occasions. G coefficients were unavailable, and D studies were not conducted. The following list reflects the findings of the study: Teacher ratings are stable over time, and teachers and observers vary substantially within a classroom.

Schroeder, Schroeder, and Hare (1983) examined the dependability of a behavioral checklist for psychopathy and a global rating of psychopathy based on Diagnostic and Statistical Manual of Mental Disorders--Third Edition (APA, 1980) as determined by two raters on two occasions. Variance component information indicated that the persons x items x raters x times + error, person, and persons x items interaction accounted for large proportions of total variance. Classical reliability analyses produced internal consistency coefficients for each rater that ranged from .82 to .92, and the interrater reliability ranged varied from .84 to .93. G coefficients ranged from .85 to .90. The results of the study indicated that the checklist is a reliable instrument with prison populations. No D study information was available.

Summary

PSYCHLIT and ERIC indicate that there are over 100 studies in the research literature since 1974 discussing GT. The priorities of those studies do not involve the evaluation of established instruments with an extensive psychometric history. Other
studies often use rating scales developed specifically for the study. Very few studies evaluate an instrument, provide information on variance components based on information collected from a G study, and discuss protocol improvement based on D study information. For a GT study to be reviewed in this section, the study must evaluate instruments with a psychometric history, include G and phi coefficients, and contain variance component information. Two GT studies of instruments that lack information about G and phi coefficients or variance components were also briefly mentioned. From the reviewed studies, the error variance component is typically one of the largest contributors to total variance.
CHAPTER III

METHODOLOGY

Sample

The data for the present study came from an extent data set collected by the Early Intervention Research Institute (EIRI) at Utah State University. EIRI was a federally funded project designed to assess the benefits and cost associated with ongoing early intervention programs. The project included more than 900 families with children identified as at risk for, or having, a developmental delay. Sixteen research sites located in nine states across the U.S. were used to investigate a specific intervention, and assessments took place approximately yearly (White, 1991; White et al., 1987). Assessment information came mainly from self-report instruments such as the PSI. Information for the present study used both control and experimental groups from the EIRI studies. The EIRI data were recoded at the item level for another federally funded project, Family Functioning in Families of Children with Disabilities: An Intensive Psychometric Investigation of Five Family Measures (Crowley, 1995). The Psychometrics Project (Crowley, 1995) represented an attempt to establish norms, reliability, and validity information on self-report, family-functioning measures gathered from families having children with disabilities.

The sample size for the present study includes 369 families whose self-report inventories were complete for three consecutive years. Demographic information for these participants is presented in Table 2 for the mothers of children with disabilities, and Table 3 contains demographic information about the children themselves. Demographic characteristics reveal that the wide range of variability of factors for mothers and children having disabilities.
### Table 2

**Mother Demographic Characteristics**

<table>
<thead>
<tr>
<th></th>
<th>Mean (SD)</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age in years</td>
<td>29.9 (6.2)</td>
<td>15.8 - 56.4</td>
</tr>
<tr>
<td>Education</td>
<td>13.6 (2.1)</td>
<td>7.0 - 17.0</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>88%</td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>9%</td>
<td></td>
</tr>
<tr>
<td>Hispanic American</td>
<td>1%</td>
<td></td>
</tr>
<tr>
<td>Native American</td>
<td>1%</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>1%</td>
<td></td>
</tr>
<tr>
<td>Marital Status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>87%</td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>7%</td>
<td></td>
</tr>
<tr>
<td>Separated</td>
<td>3%</td>
<td></td>
</tr>
<tr>
<td>Divorced</td>
<td>2%</td>
<td></td>
</tr>
<tr>
<td>Widowed</td>
<td>1%</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3

**Child Demographic Characteristics**

<table>
<thead>
<tr>
<th></th>
<th>Mean (SD)</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age in months</td>
<td>28.6 (20.4)</td>
<td>0.0 - 74.0</td>
</tr>
<tr>
<td>Type of Disability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IVH</td>
<td>28%</td>
<td></td>
</tr>
<tr>
<td>Developmental Delay</td>
<td>17%</td>
<td></td>
</tr>
<tr>
<td>Language Impaired</td>
<td>14%</td>
<td></td>
</tr>
<tr>
<td>Down Syndrome</td>
<td>7%</td>
<td></td>
</tr>
<tr>
<td>Hearing Impaired</td>
<td>6%</td>
<td></td>
</tr>
<tr>
<td>Health Impaired</td>
<td>6%</td>
<td></td>
</tr>
<tr>
<td>Multihandicapped</td>
<td>6%</td>
<td></td>
</tr>
<tr>
<td>Visually Impaired</td>
<td>6%</td>
<td></td>
</tr>
<tr>
<td>Motor Impaired</td>
<td>4%</td>
<td></td>
</tr>
<tr>
<td>Cerebral Palsy</td>
<td>3%</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>2%</td>
<td></td>
</tr>
<tr>
<td>BDI Total Score</td>
<td>141.2 (380.3)</td>
<td>0.0 - 3500.0</td>
</tr>
</tbody>
</table>
Procedures

As previously stated, self-report data were recoded at the item level from the longitudinal studies. For the Psychometrics Project at Utah State University, data were recoded using 922 subjects originally followed by a second and third year of data when available. For the present study, subjects who had completed all three years were included. The total sample size for the study included 369 subjects.

Informed Consent Procedures

The project was reviewed and approved by the Internal Review Board of Utah State University (see Appendix A). No informed consent form was necessary because archival data were being used, and the data had been recorded so that subjects could not be identified directly through participation in the study.

Measures

PSI (Long Form)

The PSI (Abidin, 1990a; see Appendix B) is a self-report inventory that attempts to assess parent/child relationships that are under stress and could possibly become child behavior problems and/or dysfunctional parent behaviors. Since its first publication in 1983, the current version of the PSI represents the third revision. The normative sample consisted of 534 parents from the first norm group in 1983, and 2,099 parents who were administered the test between 1983 and 1989. Parents were primarily volunteers from well-baby clinics recruited by agencies serving their children. The response format of the PSI is a 5-point Likert scale ranging from strongly agree to strongly disagree, and some items are reverse scored. The test contains 101 items.
divided into Parent and Child Domains. A third domain assesses life stress and is a checklist of recent stressful events. The data from the life stress scale is in a different form not appropriate for these analyses, and the life stress scale will not be the focus of the present study. Within the Parent and Child Domains, items scores are tallied to form an overall domain score, as well as subscale scores, which will be discussed in more detail shortly. The Child and Parent Domains can be combined to form a total stress score. However, some researchers have argued against the combination of scores due to the unique nature of both domains and have suggested that the examination of a total score would not add any relevant information (Boyce, Behl, Mortensen, & Akers, 1991; Innocenti et al., 1992). Therefore, the analyses for this study will focus on the Parent and Child Domains separately rather than the PSI total score.

The remainder of the present section will contain a breakdown of Child and Parent Domains. The Child Domain contains a total of 47 items. For the Child Domain, the PSI model has four subscales describing child temperament, adaptability, demandingness, mood, hyperactivity/distractibility, and two subscales, acceptability and child reinforces the parent, representing the influence of the child on the parent. The adaptability subscale contains 11 items that describe how change affects a child. "My child reacts very strongly when something happens that my child doesn't like" is an example item from the adaptability subscale. The demandingness subscale contains nine items that describe how much coercion a child uses to influence a parent to obtain something. "There are some things my child does that really bother me a lot" is an example item from the demandingness subscale. The mood subscale contains five items that describe behaviors such as crying, withdrawal, and depression. An example item from the mood subscale includes "I feel that my child is very moody and easily upset." The distractibility/hyperactivity subscale contains nine items that address areas
of activity level and on-task behavior. "My child appears disorganized and is easily distracted" is an example item from the distractibility/hyperactivity subscale. The seven-item acceptability subscale determines the proximity of the parent's ideal sense for the child and the child's actual self. An example item from the acceptability subscale includes "My child is not able to do as much as I expected." The six-item child reinforces the parent subscale entails the degree that the parent-child relationship is rewarding to the parent (Abidin, 1990a). "My child rarely does things for me that make me feel good" is an example item from the child reinforces the parent subscale.

The 54-item Parent Domain is represented by seven subscales. The depression, senses of competence, parental attachment subscales addresses the parent's personality and pathology. Four situational variables, relationship with spouse, social isolation scale, parental health, and restrictions of role, represent the four remaining Parent Domain subscales. The nine-item depression subscale describes guilt and the degree to which the parent can devote emotional resources to the child. "I feel every time my child does something wrong it is really my fault" is an example item from the depression subscale. The 13-item sense of competence subscale measures the extent of the parent's belief of being competent in their role as a parent. An example item from the sense of competence subscale includes "I feel capable and on top of things when I am caring for my child." The seven-item parental attachment subscale assesses the importance of the parent's motivation to fulfill parental responsibilities. "I expected to have closer and warmer feelings for my child than I do and this bothers me" is an example from the parental attachment subscale. The seven-item relationship with spouse and six-item social isolation subscales attempt to measure social, physical, and emotional support from the spouse and significant others, respectively. In addition, the relationship with spouse subscale also tries to estimate conflict between spouses.
because of the parenting role. An example item from the relationship with spouse subscale includes "Since having my child, my spouse (male/female friend) has not given me as much help and support as I expected." "I feel alone and without friends" is an example item from the social isolation subscale. The five-item parental health subscale attempts to evaluate how a parent's own physical needs affect their caretaking abilities. "During the past six months, I have been sicker than usual or have had more aches and pains than I normally do" is an example item from the parental health subscale. The final parent subscale, restrictions of role, contains seven items and evaluates the negative impact associated with parenthood such as losses in personal freedom and resentment (Abidin, 1990a). An example item from the restrictions of role subscale includes "I feel trapped by my responsibilities as a parent."

**PSI (Short Form)**

The theoretical model for the PSI/Short Form (PSI/SF; Abidin, 1990b; see Appendix C) is not as complex as the original PSI but retains questions from the original PSI. The PSI/SF was created to provide a brief measure of parent-child interaction that provides information about the family system. The PSI/SF consists of 36 questions taken from the original PSI and is composed of three domains, Parental Distress, Parent-Child Dysfunctional Interaction, and Difficult Child (Abidin, 1990b). Each domain contains 12 items. The Parental Distress Domain is composed mainly of items from the depression, restriction of role, social isolation, and relationship with spouse subscales from the original PSI and investigates problem areas related to the parenting role. Content areas for the Parent-Child Dysfunctional Interaction Domain include a lack of satisfaction with the child and the child's inability to attain goals expected by parents. Original items from the child acceptability, child reinforces parent, and parental
attachment subscales provide much of the basis for the Parent-Child Dysfunction Interaction domain. The final domain, Difficult Child, adopts items from the child's adaptability, demandingness, mood, and child's level of distractibility and activity subscales and attempts to address the child's ability to function independently of the parent. In a section separate from the three domains, Abidin (1990b) included a defensive responding subscale derived from the Marlow-Crowne Scale of Social Desirability to measure a parent attempting to deny existing problems in the parent-child relationship. Data from the Marlow-Crowne Scale of Social Desirability will not be analyzed due to unavailability. Twelve items compose each of the three domains, and each domain score is computed by the addition of item scores. Additionally, the three domain scores can be totaled to form the total stress score. The response format for the PSI/SF is the same as the long PSI version.

Analyses

Generalizability analysis was conducted on the data using the GENOVA program. GENOVA (generalized analysis of variance) is a FORTRAN IV system used to obtain variance estimates used in generalizability analyses (Brennan, 1983). Up to five facets can be analyzed the GENOVA program. In a G study, GENOVA is used to estimate variance components. By using the information obtained from the G study, GENOVA analyses for several alternative D studies each produce variance estimates, generalizability coefficient, and a phi coefficient as discussed in Chapter II.

For the sake of clarity, analyses will be broken down by research questions. All G studies represent a fully crossed design. Fully crossed designs represent analyses in which all facet conditions are measured across all variation sources (Shavelson & Webb, 1991a, 1991b). For example, all item conditions would be measured across all
persons conditions. For all G studies, the universe of admissible observations will contain the facets of items and occasions. These studies will consist of seven variance components that will include persons, items, occasions, persons x items interaction, persons x occasions interaction, items x occasions interaction, and a final component consisting of persons x items x occasions interaction, personal history, and error. Since three years are available, the occasions facet will consider each year as an admissible observation in all G studies. The objects of measurement for all G studies are people or more specifically families having children with disabilities.

Research question one assesses the dependability of the long PSI Parent and Child Domains. Analyses will be conducted for the Parent and Child Domains separately. The Parent Domain contains 54 items; each item will be considered as an admissible condition of the item facet. Similarly, each of the 47 items in the Child Domain will be considered as an admissible condition of the item facet analyses. The goal of the G studies is to obtain general estimates of parenting stress among families having children with disabilities using the long PSI form. For comparison purposes, internal consistency and test-retest reliability analyses were conducted on Parent and Child Domain scores using CRT.

Research question two addressed the dependability of the PSI Short Form. Although the PSI Short Form is broken down into three 12-item subscales, G study analyses will be conducted only on the total score because of the brevity of the subscales. Therefore, each of the 36 items will be considered as an admissible condition of the item facet. The goal of the G studies is to obtain general estimates of parenting stress among families having children with disabilities by using the short PSI form. For comparison purposes, internal consistency and test-retest reliability analyses will also be conducted using CRT on the PSI short form.
Question three addressed the possibility of improving dependability of the Parent Domain, Child Domain, and the short PSI form scores. D studies will investigate the generalizability and phi coefficients for multiple pairings of items and occasions for the Parent Domain, Child Domain, and the PSI short form. After examining the results, suggestions will be made for protocol and measurement improvement based upon the examination of the results and deciding the level of optimization that weighs the factors of improving reliability versus cost-effectiveness.
CHAPTER IV
RESULTS

The analysis section will be broken down by research question into G and D studies of the Parent Domain, Child Domain, and the PSI short form. The beginning of each section will focus on G study results and discuss the contribution of each facet to the total amount of shared variance. The next section will discuss the D study results. More specifically, comments about the trends in the G and phi will be discussed in detail. The D study section will also contain the comparison of the CRT analyses using internal consistency and test-retest reliability with D study results.

Parent Domain

G Study Results

The first research question addressed the dependability of the Parent Domain score across 3 years using generalizability theory. Table 4 displays the variance components for the Parent Domain. The effects of the occasions facet contributed negligibly to the percentage of total variance. Occasions as a single facet accounted for .01%, .11% for occasions x items, and 2.03% for persons x occasions. The items facet accounted for 31.60%, 23.66% for persons x occasions x items + error, and 34.58% for items, persons x items of the total variance. Items as a single facet, persons x items interaction, and persons x occasions x items + error accounted for almost 90% of the total variance.

The component of persons x occasions x items + error suggests a substantial interaction between persons, occasions, and items or a large contribution of systematic and/or unsystematic variance. Examples of systematic interaction could include
Table 4
Facets and Percentage of Total Variance in the Parent Domain

<table>
<thead>
<tr>
<th>Facet</th>
<th>Percent of Total Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>persons (p)</td>
<td>8.01</td>
</tr>
<tr>
<td>items (l)</td>
<td>31.60</td>
</tr>
<tr>
<td>occasions (o)</td>
<td>.01</td>
</tr>
<tr>
<td>persons x items (p x l)</td>
<td>34.58</td>
</tr>
<tr>
<td>persons x occasions (p x o)</td>
<td>2.03</td>
</tr>
<tr>
<td>occasions x items (o x l)</td>
<td>.11</td>
</tr>
<tr>
<td>persons x occasions x items + error</td>
<td>23.66</td>
</tr>
</tbody>
</table>

Variation in testing condition such as the time of day the test was given, the temperature of the room when the test was taken, or even something as simple as the comfort of the chair the person sat in while taking the test (Shavelson & Webb, 1991a, 1991b).

Examples of unsystematic variance could include skipping an item, answering incorrectly, or rushing through the test to get done.

The large amount of variance due to the items facet suggests that the score on a particular item does not dictate how a person would respond to other items; thus, item responses vary considerably. The interaction of persons x items suggests that the relative scores of subjects differed from item to item.

Discussion Results

The variance components from the G study are used to estimate G and phi coefficients of D studies that vary on several facet levels. For the purposes of this study, single administration of the various items, double administration of various items, and triple administration of various items will be examined. The item range
Table 5

**Parent Domain D Study Results for One, Two, and Three Administrations at Various Item Levels**

<table>
<thead>
<tr>
<th>Items, Occasions</th>
<th>Internal Consistency</th>
<th>Test-Retest</th>
<th>G Coefficient</th>
<th>Phi Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td>.89</td>
<td>.70 (1 and 2)</td>
<td>.55110</td>
<td>.48562</td>
</tr>
<tr>
<td>2 occasions</td>
<td>.89</td>
<td>.73 (2 and 3)</td>
<td>.66862</td>
<td>.57476</td>
</tr>
<tr>
<td>3 occasions</td>
<td>.89</td>
<td>.67 (1 and 3)</td>
<td>.71978</td>
<td>.61222</td>
</tr>
<tr>
<td>27 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td></td>
<td></td>
<td>.58917</td>
<td>.53231</td>
</tr>
<tr>
<td>2 occasions</td>
<td></td>
<td></td>
<td>.70717</td>
<td>.62694</td>
</tr>
<tr>
<td>3 occasions</td>
<td></td>
<td></td>
<td>.75776</td>
<td>.66643</td>
</tr>
<tr>
<td>30 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td></td>
<td></td>
<td>.60104</td>
<td>.54734</td>
</tr>
<tr>
<td>2 occasions</td>
<td></td>
<td></td>
<td>.71901</td>
<td>.64364</td>
</tr>
<tr>
<td>3 occasions</td>
<td></td>
<td></td>
<td>.76935</td>
<td>.68373</td>
</tr>
<tr>
<td>40 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td></td>
<td></td>
<td>.62956</td>
<td>.58449</td>
</tr>
<tr>
<td>2 occasions</td>
<td></td>
<td></td>
<td>.74717</td>
<td>.68466</td>
</tr>
<tr>
<td>3 occasions</td>
<td></td>
<td></td>
<td>.79679</td>
<td>.72615</td>
</tr>
<tr>
<td>54 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td></td>
<td></td>
<td>.65369</td>
<td>.61706</td>
</tr>
<tr>
<td>2 occasions</td>
<td></td>
<td></td>
<td>.77065</td>
<td>.72037</td>
</tr>
<tr>
<td>3 occasions</td>
<td></td>
<td></td>
<td>.81952</td>
<td>.76296</td>
</tr>
<tr>
<td>70 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td></td>
<td></td>
<td>.67047</td>
<td>.64036</td>
</tr>
<tr>
<td>2 occasions</td>
<td></td>
<td></td>
<td>.78879</td>
<td>.74578</td>
</tr>
<tr>
<td>3 occasions</td>
<td></td>
<td></td>
<td>.83508</td>
<td>.78908</td>
</tr>
<tr>
<td>90 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td></td>
<td></td>
<td>.68364</td>
<td>.65903</td>
</tr>
<tr>
<td>2 occasions</td>
<td></td>
<td></td>
<td>.79935</td>
<td>.76604</td>
</tr>
<tr>
<td>3 occasions</td>
<td></td>
<td></td>
<td>.84715</td>
<td>.80987</td>
</tr>
<tr>
<td>108 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td></td>
<td></td>
<td>.69156</td>
<td>.67043</td>
</tr>
<tr>
<td>2 occasions</td>
<td></td>
<td></td>
<td>.80686</td>
<td>.77838</td>
</tr>
<tr>
<td>3 occasions</td>
<td></td>
<td></td>
<td>.85435</td>
<td>.82252</td>
</tr>
<tr>
<td>120 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td></td>
<td></td>
<td>.69558</td>
<td>.67628</td>
</tr>
<tr>
<td>2 occasions</td>
<td></td>
<td></td>
<td>.81067</td>
<td>.78469</td>
</tr>
<tr>
<td>3 occasions</td>
<td></td>
<td></td>
<td>.85799</td>
<td>.82899</td>
</tr>
</tbody>
</table>

*(table continues)*
Items, Internal Occasions Test- G Phi Consistency Retest Coefficient Coefficient

<table>
<thead>
<tr>
<th>Items, Occasions</th>
<th>Internal Consistency</th>
<th>Test- Retest</th>
<th>G Coefficient</th>
<th>Phi Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>140 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td>.70083</td>
<td>.68395</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 occasions</td>
<td>.81562</td>
<td>.79297</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 occasions</td>
<td>.86273</td>
<td>.83746</td>
<td></td>
<td></td>
</tr>
<tr>
<td>150 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td>.70296</td>
<td>.68707</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 occasions</td>
<td>.81762</td>
<td>.79633</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 occasions</td>
<td>.86463</td>
<td>.84090</td>
<td></td>
<td></td>
</tr>
<tr>
<td>162 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td>.70517</td>
<td>.69033</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 occasions</td>
<td>.81970</td>
<td>.79984</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 occasions</td>
<td>.86662</td>
<td>.84450</td>
<td></td>
<td></td>
</tr>
<tr>
<td>170 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td>.70648</td>
<td>.69227</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 occasions</td>
<td>.82093</td>
<td>.80192</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 occasions</td>
<td>.86779</td>
<td>.84663</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

consisted of 20 to 170 items. Table 5 reveals that the dependability of the PSI Parent Domain score using a single administration of all 54 items yielded a G coefficient of .65 and a phi coefficient of .62 (G and phi coefficients have been rounded to the hundredths place). However, changes in the measurement protocol resulted in G coefficients ranging from .55 to .86 and phi coefficients ranging from .49 to .83. A single administration of item levels produced G coefficients steadily increasing from .55 to .71, and phi coefficients increasing from .49 to .69. A double administration of item levels produced G coefficients increasing from .69 to .82, and phi coefficients increasing from .57 to .80. A triple administration of item levels produced G coefficients increasing from .72 to .87, and phi coefficients increasing from .61 to .85.

Decreasing the number of items from 54 to 27 resulted in relatively low G and phi coefficients across all three testing occasions. Doubling or tripling the number of items (i.e., 108 or 162) across the single, double, and triple administrations predictably
increases the G and phi coefficients. For the occasions facet, the largest jump in G and phi coefficients takes place from one to two administrations. Typically, there is an increase of about .12 from Occasion 1 to Occasion 2. From Occasion 2 to Occasion 3, the increase is typically about .04 or .05. For the items facet, the increase of items is a gradual increase throughout the entire study. G study results indicated that the items facet and its interaction with other facets account for the largest share of variance. Therefore, if a researcher is interested in reducing variability, the item facet and its interaction with other facets would be worthy of investigation. One way to reduce item variability is to increase the number of items. The increase in phi and G coefficients due to more items does seem to provide an alternative for reducing item variability. However, these modifications must also consider the time and money needed to modify the PSI and the increase in time that it would take to complete additional questions.

The results reflected in Table 5 show that the dependability of PSI scores increases for G coefficients in a single administration of all 54 items (.65), double administration of 54 items (.77), and the triple administration of 54 items (.82); and for phi coefficients for a single administration of 54 items (.62), double administration of 54 items (.72), and the triple administration of 54 items (.76). An examination of these results reveals an increase of .12 from the single administration of 54 items to the double administration of 54 items. The increase in G and phi coefficients does improve substantially from Occasion 1 to Occasion 2. However, the increase from the double administration of 54 items to the triple administration of 54 items is only .04. Administering 108 items on three occasions produces an optimal G coefficient equal to .85 and a phi coefficient equal to .82. Although the triple administration of 108 items produces admirable psychometric results, it does not seem to be practical within an actual setting.
Classical reliability analyses, reported in Table 5, were also conducted to compare with G study results. Internal consistency using coefficient alpha and test-retest reliability coefficients were computed. The internal consistency results for the Parent Domain were high at .89 (Time 1), .89 (Time 2), and .89 (Time 3). Test-retest reliability analyses were in the moderate range at .70 for Time 1 and 2, .73 for Time 2 and Time 3), and .67 for Time 1 and Time 3. Classical reliability results suggest that the PSI scores from this sample have high internal consistency and moderate to high test-retest reliability. However, as shown in Table 5, GT analyses reveal that G and phi coefficient ranges are much lower. This suggests that variance component interactions lower G and phi coefficient estimates. As previously stated, CRT does not consider interactions and thus results in higher dependability estimates.

Child Domain

G Study Results

Results of Child Domain score analyses were similar to Parent Domain results. In Table 6, the variance components from the Child Domain are presented. Once again, the effects of occasions were negligible. Occasions as a single facet accounted for .01%, occasions x items for .13%, and persons x occasions for 3.17% of the total variance. The items facet accounted for 34.65%, 28.24% for persons x occasions x items + error, and 28.20% for persons x items of the total variance. Items, persons x items, and persons x occasions x items + error accounted for more than 91% of the total variance.

The large proportion of variance explained by the item facet suggests that there was substantial variability from item to item. Decreasing the items to 24 questions resulted in a substantial decrease in the dependability of test scores while increasing
Table 6
Facets and Percentage of Total Variance in the Child Domain

<table>
<thead>
<tr>
<th>Facet</th>
<th>Percent of Total Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>persons (p)</td>
<td>5.60</td>
</tr>
<tr>
<td>items (I)</td>
<td>34.65</td>
</tr>
<tr>
<td>occasions (o)</td>
<td>.01</td>
</tr>
<tr>
<td>persons x items (p x I)</td>
<td>28.20</td>
</tr>
<tr>
<td>persons x occasions (p x o)</td>
<td>3.17</td>
</tr>
<tr>
<td>occasions x items (o x I)</td>
<td>.13</td>
</tr>
<tr>
<td>persons x occasions x items + error</td>
<td>28.24</td>
</tr>
</tbody>
</table>

The items does result in an increase in G and phi coefficients. Further investigation is needed to explore the possibility of additional items to decrease variability, and thus increase dependability of test scores.

The persons x items interaction suggests that the relative standing of a person's score differed from item to item. The explanation for the persons x occasions x items + error facet could be attributed to the large interaction between persons, occasions, and items, unknown influences, or a combination of those two possibilities.

D Study Results

The dependability of the Child Domain score with a single administration produced a G coefficient equal to .42, and a phi coefficient equal to .39. As illustrated in Table 7, by increasing the number of items and administrations, the dependability of scores can be substantially improved. G coefficients ranged from .36 to .72, and phi coefficients ranged from .30 to .69. A single administration of items, ranging from 20 to 170 items, produced G coefficients steadily increasing from .35 to .47. Likewise, phi coefficients,
Table 7

Child Domain D Study Results for One, Two, and Three Administrations at Various Item Levels

<table>
<thead>
<tr>
<th>Items, Occasions</th>
<th>Internal Consistency</th>
<th>Test-Retest</th>
<th>G Coefficient</th>
<th>Phi Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>24 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td>.85</td>
<td>.61 (1 and 2)</td>
<td>.37295</td>
<td>.32491</td>
</tr>
<tr>
<td>2 occasions</td>
<td>.84</td>
<td>.68 (2 and 3)</td>
<td>.52043</td>
<td>.43151</td>
</tr>
<tr>
<td>3 occasions</td>
<td>.86</td>
<td>.58 (1 and 3)</td>
<td>.59944</td>
<td>.48449</td>
</tr>
<tr>
<td>30 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td></td>
<td></td>
<td>.39189</td>
<td>.34857</td>
</tr>
<tr>
<td>2 occasions</td>
<td></td>
<td></td>
<td>.54332</td>
<td>.46354</td>
</tr>
<tr>
<td>3 occasions</td>
<td></td>
<td></td>
<td>.62365</td>
<td>.52080</td>
</tr>
<tr>
<td>40 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td></td>
<td></td>
<td>.41285</td>
<td>.37593</td>
</tr>
<tr>
<td>2 occasions</td>
<td></td>
<td></td>
<td>.56831</td>
<td>.50071</td>
</tr>
<tr>
<td>3 occasions</td>
<td></td>
<td></td>
<td>.64989</td>
<td>.56300</td>
</tr>
<tr>
<td>47 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td></td>
<td></td>
<td>.42296</td>
<td>.38960</td>
</tr>
<tr>
<td>2 occasions</td>
<td></td>
<td></td>
<td>.58024</td>
<td>.51931</td>
</tr>
<tr>
<td>3 occasions</td>
<td></td>
<td></td>
<td>.66234</td>
<td>.58414</td>
</tr>
<tr>
<td>60 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td></td>
<td></td>
<td>.43618</td>
<td>.40797</td>
</tr>
<tr>
<td>2 occasions</td>
<td></td>
<td></td>
<td>.59572</td>
<td>.54436</td>
</tr>
<tr>
<td>3 occasions</td>
<td></td>
<td></td>
<td>.67844</td>
<td>.61283</td>
</tr>
<tr>
<td>70 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td></td>
<td></td>
<td>.44334</td>
<td>.41814</td>
</tr>
<tr>
<td>2 occasions</td>
<td></td>
<td></td>
<td>.60404</td>
<td>.55826</td>
</tr>
<tr>
<td>3 occasions</td>
<td></td>
<td></td>
<td>.68706</td>
<td>.62846</td>
</tr>
<tr>
<td>80 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td></td>
<td></td>
<td>.44887</td>
<td>.42612</td>
</tr>
<tr>
<td>2 occasions</td>
<td></td>
<td></td>
<td>.61044</td>
<td>.56917</td>
</tr>
<tr>
<td>3 occasions</td>
<td></td>
<td></td>
<td>.69367</td>
<td>.64088</td>
</tr>
<tr>
<td>94 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td></td>
<td></td>
<td>.45478</td>
<td>.43476</td>
</tr>
<tr>
<td>2 occasions</td>
<td></td>
<td></td>
<td>.61725</td>
<td>.58100</td>
</tr>
<tr>
<td>3 occasions</td>
<td></td>
<td></td>
<td>.70070</td>
<td>.65436</td>
</tr>
</tbody>
</table>

(table continues)
Items, Occasions | Internal Consistency | Test-Retest | G Coefficient | Phi Coefficient
--- | --- | --- | --- | ---
100 items, 1 occasion | | | .45684 | .43781
2 occasions | | | .61963 | .58517
3 occasions | | | .70314 | .65912
120 items, 1 occasion | | | .46231 | .44596
2 occasions | | | .62590 | .59634
3 occasions | | | .70960 | .67186
141 items, 1 occasion | | | .46647 | .45224
2 occasions | | | .63006 | .60475
3 occasions | | | .71449 | .68168
160 items, 1 occasion | | | .46934 | .45660
2 occasions | | | .63393 | .61093
3 occasions | | | .71785 | .68851
180 items, 1 occasion | | | .47173 | .46025
2 occasions | | | .63666 | .61595
3 occasions | | | .72064 | .69424

ranging from 20 to 170 items, increased from .30 to .46 for a single administration. A double administration of item levels produced G coefficients increasing from .50 to .64, and phi coefficients increased from .40 to .62. A triple administration of item levels produced G coefficients increasing from .58 to .72, and phi coefficients increasing from .45 to .69.

Similar to Parent Domain results, reducing the number of items to 20 produces very low G and phi coefficients. Although increasing the number of items across multiple administrations increases G and phi coefficients, the results are still low. For the occasions facet, the largest increase of coefficients is accounted for by the increase from Occasion 1 to Occasion 2, which is approximately equal to .15. The increase from
Occasion 2 to Occasion 3 is not as large but is approximately equal to .08. The increase in the number of administered items yields a gradual increase. Since the item facet and its interaction with other facets represent the largest proportion of total variance, improvements in the Child Domain should focus on item variability. The results of the Child Domain are so inconsistent that three administrations of 180 items produce a G coefficient equal to .72, and a phi coefficient equal to .69. An acceptable range for reliability coefficients begins at approximately .80 (Shavelson & Webb, 1991a, 1991b). These low results suggest that the Child Domain has lower consistency in comparison to the Parent Domain results.

The internal consistency analyses were similar to the results reported for the Parent Domain at .85 (Time 1), .84 (Time 2), and .86 (Time 3). However, the test-retest reliability coefficients for the Child Domain at .61 for Time 1 and Time 2, .68 for Time 2 and Time 3, and .58 for Time 1 and Time 3 were somewhat lower than Parent Domain scores. By using all 47 Child Domain items, a single administration (.42), a double administration (.58), and a triple administration (.66) reflect the increase in the G coefficient across the various test occasions. Phi coefficient values also produce an increase across single (.39), double (.52), and triple administrations (.58). GT analyses contained lower ranges than the CRT ranges possibly resulting from the variable interaction that CRT ignores.

Short Form

G Study Results

The variance components are presented in Table 8. As in previous GT analyses, occasion as a single facet or combined with other facets contributed minimally. Occasions as a single facet contributed .01%, occasions x items accounted for .18%,
Table 8

Facets and Percentage of Total Variance in the Short PSI

<table>
<thead>
<tr>
<th>Facet</th>
<th>Percent of Total Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>persons (p)</td>
<td>10.47</td>
</tr>
<tr>
<td>items (I)</td>
<td>29.99</td>
</tr>
<tr>
<td>occasions (o)</td>
<td>.01</td>
</tr>
<tr>
<td>persons x items (p x I)</td>
<td>31.93</td>
</tr>
<tr>
<td>persons x occasions (p x o)</td>
<td>3.05</td>
</tr>
<tr>
<td>occasions x items (o x I)</td>
<td>.18</td>
</tr>
<tr>
<td>persons x occasions x items + error</td>
<td>24.36</td>
</tr>
</tbody>
</table>

and persons x occasions accounted for 3.05% of the total variance. The facets of persons x occasions x items + error accounted for 24.36% of the total variance, items explained 29.99% of the total variance, and persons x items accounted for 31.93% of the total variance. Persons x occasions x items + error, persons x items, and items accounted for more than 86% of the total variance.

Possible explanations for the contribution of the persons x occasions x items + error variance component include error interaction, systematic variance, or unsystematic variance. The contribution of the items variance component is related to the idea that a score on a particular item does not predict how a person would respond to other items; thus, item responses vary considerably. The large contribution of the interaction of persons x items variance components would indicate that the subject scores differed from item to item.
D Study Results

A single administration of the PSI Short Form using all 36 items yielded a G coefficient of .62 and a phi coefficient of .58. However, as shown in Table 9, changes in the measurement protocol resulted in G coefficients ranging from .55 to .86 and phi coefficients ranging from .50 to .84. A single administration of items, ranging from 18 to 160 items, produced G coefficients steadily increasing from .55 to .68. Likewise, phi coefficients, ranging from 18 to 160 items, increased from .50 to .67. The range of G and phi coefficients represents a low range of consistency.

The increase from Occasion 1 to Occasion 2 represents the largest jump of approximately .12 in coefficients when looking at the number of administrations. The increase from Occasion 2 to Occasion 3 is approximately equal to .05. The coefficients increase for the items facet is gradual over the D studies without any large increases. Decreasing the number of items from 36 to 18 resulted in relatively low G and phi coefficients across all three administrations. The range of G coefficients for 18 items across all administrations ranged from .55 to .74, and the range of phi coefficients for 18 items across all administrations ranged from .50 to .64. Doubling the number of items to 72 questions across all three administrations resulted in G coefficients ranging from .66 to .84 and phi coefficients ranging from .64 to .80. Tripling the number of items to 108 across the various administrations produced G coefficients ranging from .67 to .85, and phi coefficients ranging from .66 to .83. Although coefficients did increase over occasions, the effects of time were negligible, and the facet of items and its interactions with other variance components yielded the most information about total variability.

Classical reliability analyses were conducted using internal consistency using coefficient alpha and test-retest reliability coefficients. Internal consistency results were fairly high at .88 (Time 1), .88 (Time 2), and .88 (Time 3). Test-retest reliability analyses
Table 9
Short PSI D Study Results for One, Two, and Three Administrations at Various Item Levels

<table>
<thead>
<tr>
<th>Items, Occasions</th>
<th>Internal Consistency</th>
<th>Test-Retest</th>
<th>G Coefficient</th>
<th>Phi Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>18 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td>.88</td>
<td>.65 (1 and 2)</td>
<td>.55433</td>
<td>.49797</td>
</tr>
<tr>
<td>2 occasions</td>
<td>.88</td>
<td>.70 (2 and 3)</td>
<td>.68061</td>
<td>.59781</td>
</tr>
<tr>
<td>3 occasions</td>
<td>.88</td>
<td>.78 (1 and 3)</td>
<td>.73653</td>
<td>.64062</td>
</tr>
<tr>
<td>30 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td></td>
<td></td>
<td>.60585</td>
<td>.56397</td>
</tr>
<tr>
<td>2 occasions</td>
<td></td>
<td></td>
<td>.73225</td>
<td>.67214</td>
</tr>
<tr>
<td>3 occasions</td>
<td></td>
<td></td>
<td>.78697</td>
<td>.71804</td>
</tr>
<tr>
<td>36 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td></td>
<td></td>
<td>.62027</td>
<td>.58330</td>
</tr>
<tr>
<td>2 occasions</td>
<td></td>
<td></td>
<td>.74641</td>
<td>.69370</td>
</tr>
<tr>
<td>3 occasions</td>
<td></td>
<td></td>
<td>.80068</td>
<td>.74042</td>
</tr>
<tr>
<td>50 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td></td>
<td></td>
<td>.64164</td>
<td>.61270</td>
</tr>
<tr>
<td>2 occasions</td>
<td></td>
<td></td>
<td>.76717</td>
<td>.72633</td>
</tr>
<tr>
<td>3 occasions</td>
<td></td>
<td></td>
<td>.82070</td>
<td>.77418</td>
</tr>
<tr>
<td>60 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td></td>
<td></td>
<td>.65125</td>
<td>.62623</td>
</tr>
<tr>
<td>2 occasions</td>
<td></td>
<td></td>
<td>.77643</td>
<td>.74127</td>
</tr>
<tr>
<td>3 occasions</td>
<td></td>
<td></td>
<td>.82959</td>
<td>.78962</td>
</tr>
<tr>
<td>72 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td></td>
<td></td>
<td>.65949</td>
<td>.63796</td>
</tr>
<tr>
<td>2 occasions</td>
<td></td>
<td></td>
<td>.78432</td>
<td>.75419</td>
</tr>
<tr>
<td>3 occasions</td>
<td></td>
<td></td>
<td>.83714</td>
<td>.80296</td>
</tr>
<tr>
<td>90 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td></td>
<td></td>
<td>.66793</td>
<td>.65015</td>
</tr>
<tr>
<td>2 occasions</td>
<td></td>
<td></td>
<td>.79237</td>
<td>.76758</td>
</tr>
<tr>
<td>3 occasions</td>
<td></td>
<td></td>
<td>.84097</td>
<td>.80980</td>
</tr>
<tr>
<td>108 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td></td>
<td></td>
<td>.67369</td>
<td>.65853</td>
</tr>
<tr>
<td>2 occasions</td>
<td></td>
<td></td>
<td>.79783</td>
<td>.77677</td>
</tr>
<tr>
<td>3 occasions</td>
<td></td>
<td></td>
<td>.85004</td>
<td>.82622</td>
</tr>
</tbody>
</table>

(table continues)
were in the low range at .65 for Time 1 and Time 2, .70 for Time 2 and Time 3, and .78 for Time 1 and Time 3. Classical reliability results suggest that the PSI short form scores from this sample have high internal consistency and moderate test-retest reliability. The results reflected in Table 9 indicate that the dependability of PSI scores increases over testing occasions using all 36 test items for G coefficients for one (.62), two (.75), and three (.80) administrations. Following a similar trend, phi coefficients increased from one (.58), two (.69), and three (.74) administrations for all 36 items. Once again, GT analyses produced lower G and phi coefficients in Table 9 than the results of CRT analyses. CRT analyses will often produce higher results because variable interactions are not considered.

<table>
<thead>
<tr>
<th>Items, Occasions</th>
<th>Internal Consistency</th>
<th>Test-Retest</th>
<th>G Coefficient</th>
<th>Phi Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>120 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td>.67660</td>
<td>.66281</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 occasions</td>
<td>.80059</td>
<td>.78145</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 occasions</td>
<td>.85267</td>
<td>.83104</td>
<td></td>
<td></td>
</tr>
<tr>
<td>130 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td>.67863</td>
<td>.66580</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 occasions</td>
<td>.80251</td>
<td>.78472</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 occasions</td>
<td>.85450</td>
<td>.83440</td>
<td></td>
<td></td>
</tr>
<tr>
<td>144 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td>.68102</td>
<td>.66932</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 occasions</td>
<td>.80476</td>
<td>.78857</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 occasions</td>
<td>.85664</td>
<td>.83836</td>
<td></td>
<td></td>
</tr>
<tr>
<td>150 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td>.68191</td>
<td>.67064</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 occasions</td>
<td>.80560</td>
<td>.79002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 occasions</td>
<td>.85744</td>
<td>.83985</td>
<td></td>
<td></td>
</tr>
<tr>
<td>160 items,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 occasion</td>
<td>.68325</td>
<td>.67263</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 occasions</td>
<td>.80686</td>
<td>.79219</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 occasions</td>
<td>.85864</td>
<td>.84208</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER V
DISCUSSION

The major questions of this study addressed the dependability of the Parent and Child Domains of the long PSI form and the total score of the short PSI form. In addition, the final research question asked what modifications researchers could make in the number of administered items and test occasions to improve the dependability of the Parent Domain in the long PSI form, Child Domain in the long PSI form, and the total score on the short form. The following section will begin with a review of the results followed by the interpretations of the results, limitations of the study, and suggestions for future research. The implication of the results will be continued in more detail from their brief mention in the analyses section. Since the results from the G studies are similar across the Parent Domain, Child Domain, and total score from the short version, discussion of G and D studies will encompass the results from all three research areas.

Review and Interpretation of the Results

G Studies

The results for the three areas of interest, Parent and Child Domains, and PSI short form total score, varied only slightly. Overall, the items facet contributed large percentages to the total variance component, and the occasions facet contributed only small amounts to the total variance component. The numerical contributions of the variance components for the Parent Domain and the short PSI were similar. This is an interesting result since only 15 (42%) of the 36 items on the PSI short form come from the Parent Domain in the long PSI version. The speculation for the reason of this finding cannot be made with any certainty; however, one possible suggestion is that the
Parent Domain and PSI short form total scores provide a more similar portrayal of families having children with disabilities. The order and numerical contributions of the variance components for the Child Domain differed from Parent Domain and the PSI short form total score results. However, despite these slight differences, the items facet and its interaction with other variance components shared the most variance with the total variance component. Occasions and its interactions with other variance components only made small contributions to the total variance component. Although the persons variance component contributed more than the occasions facet and its interactions, the contribution of the persons component in the Child Domain could also be described as minimal.

In each case, the items facet and its interaction with other variance components shared the most variance with the total variance. The contributions of the items facet in the G and D studies are unlike several reviewed GT studies in Chapter II. Typically, consistency across items is a common finding in the reviewed GT studies with other instruments. The prominence of the items facet alone dictates a wide range of variability in how respondents answered items. The PSI answer format is arranged on a scale from strongly agree to strongly disagree. The implication for the great contribution of item variability is that a response pattern does not appear for subjects. Past stress studies (Hanson & Hanline, 1990; Kazak & Marvin, 1984) indicate that families having children with disabilities report higher stress levels than families not having children with disabilities, and stress levels for families having children with disabilities is stable over time. Results from this study indicate that those stress levels seem to vary, potentially due to different forms of stress faced by families having children with disabilities. Developmentally, family and child needs change over time. Therefore, the item
variability in this study may reflect the growing and changing needs of the family and the child with a disability.

The interaction of persons x items suggests that a person’s relative standing changed from item to item. For example, this result suggests that a person who responded strongly agree to one item does not necessarily respond to other items with the "strongly agree" response. In the review of GT studies, two of the three studies using items as a facet reported that the persons x items variance component was one of the top contributors to the total variance component. Therefore, a similarity between this study and other reviewed GT studies is the prominence of the persons x items variance component.

Several explanations may account for item variability. One possible solution for the large contribution of the items facet and its interaction with other variance components to the total variance component may be due to item diversity. The Parent Domain contains seven subscales that investigate different interest areas such as parental health, depression, attachment, role restrictions, sense of competence, social isolation, and relationship with spouse. The Child Domain contains six subscales that examine areas such as adaptability, acceptability, demandingness, mood, distractibility/hyperactivity, and reinforcement of parent. The PSI short form represents a conglomeration of questions from the Parent and Child Domains. The areas addressed by the Parent Domain, Child Domain, and the short form may represent such diverse areas that the large proportion of variability due to the items facet and its interaction with other variance components is simply due to divergent content in subscales.

To explore this possibility, G and D studies were conducted on a single subscale for each area of interest. The sense of competence scale was selected from the Parent Domain, the adaptability subscale was selected from the Child Domain, and the parent-
child interaction subscale was selected from the PSI short form. Results of the G studies revealed that the contributions of the items facet and its interaction with other variance components reflected the same trend established by the examination of all subscales in the Parent Domain, Child Domain, and PSI short form versions. Items, persons x items, and persons x items x occasions + error made the largest contributions to the total variance component. Occasions, persons x occasions, occasions x items, and persons made small to moderate contributions to the total variance component.

The results of the subscale analyses suggest that even the subscales have a wide range of variability in the PSI short and long forms. The multifaceted nature of the PSI forms may be responsible for item variability. Instead of measuring a single concept, PSI subscales attempt to measure several different areas. For example, the Parent Domain sense of competence subscale may address several issues such as self-efficacy and parental responsibilities. For example, items from the sense of competence subscale such as "I can't make decisions without help" and "I often have the feeling that I cannot handle things very well" reflect self-efficacy issues. Yet other items, such as "I have had many more problems raising children than I expected" and "My child seems to be much harder to care for than most," more clearly reflect attitudes and beliefs about parental duties. A 13-item subscale may not completely assess these domains. Therefore, parents may score higher or lower in certain areas, thus producing increased item variability. Thus, PSI questions may not thoroughly or consistently assess specific domains such as parental competence, depression, and so forth, addressed by the instrument. However, previous research results and the strong psychometric history of the PSI would suggest that it does provide an accurate overall level of parental stress. Therefore, the PSI may not be a good measure of specific
problem areas, but may provide accurate information about the overall stress level of a parent due to parent/child interactions.

The persons variance component indicated that parents somewhat differed in their perception of parental stress. The implication of this finding suggests that variability exists within the test scores of this study's population of families having children with disabilities. The perception of parental stress is different from parent to parent.

The effects of occasions and its interaction with other variance components produced inconsequential amounts of shared variance with the total variance. This finding suggests that PSI scores for families having children with disabilities are consistent across occasions. The occasions facet may have been affected by the time interval between data collections. The data collection in this study took place over intervals of at least a year. In most studies, data are collected on shorter intervals such as a few weeks or months. When the data collection interval is long, instrument and behavioral instability combine, which ultimately reduces the occasions variance component. Therefore, the effect of occasions is probably underestimated in this study due to the long time periods between data collections.

The error component suggests that a substantial amount of variance has been explained due to an interaction of the variance components (systematic variance), sources of parenting stress that were not captured by the PSI such as financial difficulties, unemployment, and so forth (unsystematic variance), or a combination of both alternatives. A common finding among GT studies is the large contribution of the error component. Even with multi-examination of error sources such as occasions, items, or raters in GT, the common error component reminds researchers that many known and unknown variables decrease test score dependability.
The trend for all of the D studies for the Parent Domain, Child Domain, and PSI short form total score indicates that G and phi coefficients increase with elevations in occasions and items. For the occasions facet, the big increase in G and phi coefficients took place in an increase with one to two administrations. The increase from Occasion 1 to Occasion 2 for the Parent Domain, Child Domain, and PSI short form was at least .12 for G and phi coefficients. The increase of administered items produced a gradual increase in G and phi coefficients for the Parent Domain, Child Domain, and the PSI short form total score. Typically, reliability coefficients are considered within an acceptable range at approximately .80 (Shavelson & Webb, 1991a, 1991b). Therefore, the optimal level of occasions and items will be when the G coefficient reaches the .80 level. It should be pointed out that the .80 range is also consistent with CRT analyses. This is a very rigorous standard for GT analyses because more error sources are considered than in CT analyses. Though the .80 level will be used in this study, it should be acknowledged that this standard is very rigorous by most psychometric standards. A single administration of the 47 items in the Child Domain, 54 items in the Parent Domain, and the 36 items of the PSI short form produces G and phi coefficients that researchers would classify in a low range of dependability ranging from .42 to .65. For the Parent Domain, two administrations of 100 items are equal to .85. Since the Child Domain results did not reach the .80 level, additional analyses were conducted. Results from the Child Domain reach the .80 level with eight administrations of 47 items. Therefore, eight administrations of 47 items represent the optimal level for the Child Domain that results in a G coefficient equal to .80. The optimal level for the total score for the short PSI version was equal to three administrations of 36 items, producing a G coefficient equal to .80. Despite the appealing psychometric characteristics of optimal
item and occasion administrations, the practical appeal to the PSI of these standards represents an unreasonable application for practitioners.

The trend for all of the D studies, including the subscale analyses, indicates that G and phi coefficients increase with multiple administrations and the elevation of items. Therefore, the best G and phi coefficients are produced with three PSI administrations as many as three times the number of items in each scale. However, three PSI administrations in a single setting with the optimum number of items produced in the D studies would require a subject to fill out as many as 540 items and is impractical. Continuing to increase administration and item levels would produce impressive psychometric characteristics. However, in an early intervention or other practical settings, increased item and administration levels would be impractical and intolerable for subjects. The optimal number of administrations and items will be discussed within a practical setting in the protocol improvement section.

In order to choose between G and phi coefficient interpretive information, past evidence must be reviewed. Families having children with disabilities have typically reported higher stress levels than families who do not have children with disabilities (Hanson & Hanline, 1990; Kazak & Marvin, 1984). The mean total stress score for Abidin's (1990a) long PSI standardization sample was equal to 222, and Crowley's (1995) mean total stress score for the PSI on families having children with disabilities was equal to 333. In the original short version standardization sample, Abidin (1990b) reported that the mean total score for the standardized group was equal to 71.0, but a study examining the norms for parents having children with disabilities on the PSI short form found the mean total score to be equal to 82.6 (Smith & Innocenti, 1993). If families having children with disabilities have higher stress levels than families not having children with disabilities, the cutoff scores for families having children with
disabilities should be higher than the cutoff scores for families not having children with disabilities.

Abidin (1990a) recommended that parents whose long PSI score is above 260 or whose short PSI score is above 90 should be identified as at risk for dysfunctional parental behaviors or child behavioral problems. Those cutoff levels were established for families having children without disabilities. PSI cutoff scores for families having children with disabilities are unavailable. Abidin (1990b) stated that PSI scores of families having children with disabilities may differ from PSI norms, and professionals should be aware of unique issues faced by families having children with special needs. Norms for some groups such as attention deficit disorder, autism, cerebral palsy, child abuse, developmentally delayed, and hyperactive children are available in the PSI manual (Abidin, 1990a). The profiles are based on very small samples of those groups ranging from 10 to 20 parents for each sample, and cutoff scores are not included for these groups.

G or phi coefficients represent dependability estimates in GT analyses. Researchers must choose which coefficient accurately describes test scores. For screening purposes, researchers and practitioners often use a cutoff score for identification of problem areas. When cutoff scores are available, phi coefficients are used to interpret absolute performance. However, G coefficients provide the best interpretation for the PSI due to the following reasons: Families having children with disabilities report having higher stress scores than families having children without disabilities; PSI profiles of families having children with special needs are very small samples and thus could be unrepresentative; few groups represent families having children with disabilities; and cutoff scores are unavailable even for those families having children with special needs.
Classical Reliability Theory and Generalizability Theory Analyses

Traditional classical reliability analysis studies have shown that the PSI forms have performed well (Abidin, 1990a, 1990b). Internal consistency and test-retest reliability information were conducted to compare with information gained from G and D studies. Overall, internal consistency information was in the high range, .88 to .89. Although test-retest reliability information was not as high as internal consistency analyses, the time interval may have had a large impact on the test-retest analyses. Most test-retest analyses are conducted over a short interval such as a few weeks or a month. The time interval for the test-retest analyses in this study was at least a year. Although the test-retest information seems low in comparison to other reliability information, test-retest analyses in this study would be considered in the average to the high range. Scores from test-retest reliability information over the first and second years for the Parent Domain, Child Domain, and PSI short form total score extended from .58 to .78.

Another important factor to consider in the discussion of stress is fluctuating scores. Stress scores are largely impacted by daily and weekly hassles. If stress scores fluctuate in short intervals, just as much fluctuation could be expected in intervals over extended time periods. GT studies by Schroeder et al. (1983) and Crowley et al. (1994) also included information on CRT analyses. Internal consistency data collected in both studies produced high coefficients similar to the internal consistency information from this study, and the test-retest reliability results were similar to CRT analyses conducted in a study by Crowley et al. (1994).

Results from the CRT analyses indicate that internal consistency is high, but test-retest reliability is in the average range. Eason (1991) pointed out results from GT and CRT analyses can conflict, and the results from this study reinforce that statement.
Internal consistency procedures examine item variability, and low reliability suggests that test items measure low consistency of examinee responses (Crocker & Algina, 1986). According to CRT analyses, variability in PSI scores is best explained by the effects of occasions rather than the contribution of item variability.

GT analyses examine multiple error sources, and thus, typically produce lower results than CRT analyses that only examine one error source at a time. GT analyses suggest that not only is the occasions facet a negligible factor but so are the interactions of occasions and other variance components. Results of the G studies indicate that the items facet and its interaction with other variance components are responsible for the majority of variability within the study. CRT analyses cannot break down the contribution of error sources into individual components such as items, occasions, and various interactions (Crocker & Algina, 1986; Traub, 1994). GT can separate sources of error and the interaction of facets in a single analysis (Shavelson & Webb, 1991a, 1991b). CRT may incorrectly attribute variability to false sources (Eason, 1991), similar to the results in this study.

The large contribution of the items facet indicates that there may not be any established trends in item responses. The low contribution of the occasions facet indicates multiple administrations of items do not play a large role except in the coefficient increase from Occasion 1 to Occasion 2. However, this interpretation should be made with caution because the occasions facet may be underestimated due to the long time period between data collections.

One explanation for item variability suggests that the PSI is an ineffective measure, but previous psychometric information collected on the PSI discounts this hypothesis. A more plausible alternative indicates that the PSI forms are an inadequate measure for specific information such as depression, sense of competence, and so forth, provided
by subscales, but the PSI forms provide an accurate portrayal of an overall stress level due to parent-child information.

D study analyses indicate that the highest G and phi coefficients are produced with the highest number of administrations and items. However, administering the highest number of administrations and items would be impractical within any setting. In most cases with cutoff scores, phi coefficients provide the most information about absolute decisions. G coefficients will be used in this study to describe the rank comparison of individuals since cutoff scores for families having children with disabilities are unavailable.

CRT analyses indicate that the occasions component is responsible for the most variability, and item variability is quite low. However, GT analyses suggest that the items facet represents a larger influence than the occasions facet. These contradictory results between GT and CRT analyses often occur probably because of GT's multiple error source analyses compared to CRT's examination of a single error source in one analysis.

Protocol Improvement

The final question proposed in the study asked what alternatives could be made to improve the dependability of the Parent Domain, Child Domain, and the short version of the PSI. The results of both the G and D studies provide clues for protocol improvement. D study analyses suggest that multiple administrations should be conducted for all three areas. However, the contribution of the occasions facet and its interactions was often represented by negligible amounts of the total variance. With the exception of the increase in G coefficients from Occasion 1 to Occasion 2, administrations beyond two would increase G coefficients negligibly. The items facet
and its interaction with other variance components represented the largest contributors to the total variance component. Therefore, the manipulation of item levels for the three areas illustrates the best available method to improve the dependability of test scores. As evident in the D studies, decreasing the number of administered items produces lower G and phi coefficients. Consequently, increasing the number of items in each of the three areas increases the G and phi coefficients. With the increase in G and phi coefficients from Occasion 1 to Occasion 2, a double administration of more items might be the most effective method to expand score dependability.

For the Parent Domain, an alternative for protocol improvement would mean the addition of items. If one were to consider only psychometric characteristics and targeting a G coefficient equal to .80, the Parent Domain should have 100 administered twice or 200 items in a single setting. For the Child Domain, three administrations of 180 items or 540 items in a single setting should be given. The total score for the PSI short form would reach acceptable levels with three administrations of 36 items or the single administration of 108. However, these recommendations are obviously impractical because the PSI long form would be 740 items long, and the PSI short form would be slightly longer than the current existence of the PSI long form.

A more practical set of recommendations should be extended for practitioners and researchers. For the 54-item Parent Domain and the 36-item PSI short form total score, administering the original number of items twice results in G coefficients in a moderate range. Although the resulting G coefficients do not reach the acceptable range of .80, the G coefficient for the Parent Domain is equal to .77, and the G coefficient for the short PSI total score is equal to .75. Administering the original number of items twice represents a balance between improving psychometric characteristics and practical use for the PSI. Recommendations for the Child Domain are difficult to make since G
coefficients do not reach a moderate range of dependability until six administrations of 24 items. Although administering the original number of items twice only produces a G coefficient equal to .58, increasing item and administration levels beyond this point infringes on the practical limitations for researchers and practitioners. Thus, the 54-item Child Domain should also be administered twice. These results should be used with some caution, since the scores have only moderate dependability. In summary, the original number of items from the Parent Domain, Child Domain, and PSI short form total score should be administered twice to increase the dependability of scores and still fall within practical limitations.

Limitations

Caveats and limitations exist in every study, and this study is no exception. Three prominent limitations will be discussed specific to the study's results. Overall, caution should be issued to researchers and practitioners attempting to generalize the results of this study to other groups taking the PSI. First, the collected data represented 3 years of evaluation for families having children with a disability. The EIRI studies began with 922 families participating in early intervention assessment, and the original EIRI longitudinal studies had an attrition rate of less than 10% (White, 1991). The information for this study came from extant longitudinal data that were not designed to be combined across sites. Each of the EIRI studies represented a comparative study, and each study examined different study variables such as type of disability and age of children and families (White, 1991). Families completed assessment information at different points as a result of the design of the longitudinal analyses. Therefore, the design of the EIRI analyses had 922 families complete the PSI at the initial assessment, but data were available for only 369 families for the analyses in this study. Information
regarding why families dropped out of the project is unknown, and the effects of the attrition rates on the analyses are unknown. (The rationale for this difference varies by each specific site and the impact of these changes in test administration on the present analyses is unknown.)

Several cautions should be extended regarding the subjects from this study. The results of these analyses should only be compared to families having children with disabilities. Unless a similar study is performed on families having children without disabilities, comparisons of these analyses should only be applied to families having children with disabilities. Furthermore, the subjects were volunteers, and the results from these analyses may differ from nonvolunteers having children with disabilities (Anastasi, 1988). Also, the primary care giver of a child with a disability answered the PSI questions in the EIRI study. Typically, the primary care giver was the child's mother. Past studies examining families having children with disabilities have found differences between mothers and fathers having children with a disability (Beckman, 1991; Bristol, Gallagher, & Schopler, 1988; Goldberg, Morris, Simmons, Fowler, & Levison, 1990; Krauss, 1993; Tavormina, Boll, Dunn, Luscomb, & Taylor, 1981). The analyses performed in this study generalize to volunteer mothers of a children with disabilities.

Information gathered from the PSI is specific material regarding the nature of the parent/child relationship from families having children with disabilities. The PSI represents a parental stress instrument and not a generic life stress inventory. Comparisons to other studies examining stress variables not associated with parental stress factors should be made with extreme caution.
Future Recommendations

The underlying question of this study asks if the PSI should be used with families having children with disabilities. Attempts have been made to establish norms for the PSI long and short versions for families having children with disabilities (Crowley, 1995; Innocenti et al., 1992; Smith & Innocenti, 1993). In order for the PSI versions to have practical applications for professionals, cutoff scores must be established for families having children with disabilities. Cutoff scores help professionals determine if a family requires outside assistance due to the comparison of their score to a similar group norm.

The recommendations for the PSI versions are dependent upon their future use. If administrators, researchers, and practitioners wish to use the PSI simply as a measure of total parental stress based on parent and child characteristics, the PSI long and short versions do an adequate job of performing that task based on the results of this study. Some professionals want to use the PSI as a more informative source of parent and child characteristics by interpreting subscale scores. Yet, the analyses from this study suggested that item variability was largely responsible for total variance. Even the examination of largest subscales in the parent, child, and short version areas revealed that item variability continued to plague the PSI versions. Due to the nature of item variability, subscale breakdowns would produce interpretations that are likely to be unstable and yield undependable data. Many other researchers have questioned using the PSI total score (Innocenti et al., 1992) due to its multifaceted nature. The results of this study suggest that interpretation at the subscale level is also questionable. Abidin (1990a) recommended that PSI subscales should not be interpreted. Therefore, the results of these analyses concur with Abidin's (1990a) recommendation that subscale
interpretation should not be done for families having children with disabilities. If professionals wish to use the PSI versions for this purpose, the PSI forms should be altered to decrease item variability. One alternative would be to add more items to the subscales to improve item variability.

A researcher or practitioner may want information to decide what form, long or short, to choose. The correlation between the total stress score of the full-length PSI and the total stress score of the short PSI is .94 (Abidin, 1990b). The results from this study suggest the PSI short version has psychometric properties equal to or better than the long form domains. On the basis of this psychometric study, the short PSI form would be recommended. However, in choosing to use the short form, some information will not be accounted for in the short PSI version that is accounted for by the long PSI version. Researchers or practitioners should make their choice on the desired amount of pertinent information. If the PSI is to be used as a quick screening tool or as one test in a complete assessment, the short form may be of more use. If the PSI is to be used as a primary source of information about parent and child interactive systems, the long PSI version would be recommended.
REFERENCES


APPENDICES
Appendix A:

Institutional Review Board Approval
November 15, 1995

TO: Susan L. Crowley - PI
    Jim D. Sharpnack - Student Researcher

FROM: True Rubal, IRB Secretary

SUBJECT: Proposal Titled, "An Investigation of the Parenting Stress Index in the Context of Generalizability Theory"

The above-referenced proposal has been reviewed by this office and is exempt from further review by the Institutional Review Board. The IRB appreciates researchers who recognize the importance of ethical research conduct. While your research project does not require a signed informed consent, you should consider (a) offering a general introduction to your research goals, and (b) informing, in writing or through oral presentation, each participant as to the rights of the subject to confidentiality, privacy or withdrawal at any time from the research activities.

The research activities listed below are exempt from IRB review based on the Department of Health and Human Services (DHHS) regulations for the protection of human research subjects, 45 CFR Part 46, as amended to include provisions of the Federal Policy for the Protection of Human Subjects, June 18, 1991.

1. Research conducted in established or commonly accepted educational settings, involving normal educational practices, such as (a) research on regular and special education instructional strategies, or (b) research on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods.

2. Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless: (a) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and (b) any disclosure of the human subjects' responses outside the research could
reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.

3. Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior that is not exempt under paragraph (2)(b) of this section, if: (a) the human subjects are elected or appointed public officials or candidates for public office; or (b) federal statute(s) require(s) without exception that the confidentiality of the personally identifiable information will be maintained throughout the research and thereafter.

4. Research, involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.

5. Research and demonstration projects which are conducted by or subject to the approval of department or agency heads, and which are designed to study, evaluate, or otherwise examine: (a) public benefit or service programs; (b) procedures for obtaining benefits or services under those programs; (c) possible changes in or alternatives to those programs or procedures; or (d) possible changes in methods or levels of payment or benefits or services under those programs.

6. Taste and food quality evaluation and consumer acceptance studies, (a) if wholesome foods without additives are consumed, or (b) if a food is consumed that contains a food ingredient at or below the level and for a use found to be safe, or agricultural chemical or environmental contaminant at or below the level found to be safe, by the Food and Drug Administration or approved by the Environmental Protection Agency or the Food Safety and Inspection Service of the U.S. Department of Agriculture.

Your research is exempt from further review based on exemption number 4. Please keep the committee advised of any changes, adverse reactions or termination of the study. A yearly review is required of all proposals submitted to the IRB. We request that you advise us when this project is completed, otherwise we will contact you in November of 1996.
Appendix B:

Parenting Stress Index--Long Form
PARENTING STRESS INDEX (PSI)

Administration Booklet

Richard R. Abidin
Institute of Clinical Psychology
University of Virginia

Directions:

In answering the following questions, please think about the child you are most concerned about.

The questions on the following pages ask you to mark an answer which best describes your feelings. While you may not find an answer which exactly states your feelings, please mark the answer which comes closest to describing how you feel. **YOUR FIRST REACTION TO EACH QUESTION SHOULD BE YOUR ANSWER.**

Please mark the degree to which you agree or disagree with the following statements by filling in the number which best matches how you feel. If you are not sure, please fill in #3.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly Agree</td>
<td>Agree</td>
<td>Not Sure</td>
<td>Disagree</td>
<td>Strongly Disagree</td>
</tr>
</tbody>
</table>

*Example: 1 2 3 4 5*  I enjoy going to the movies. (If you sometimes enjoy going to the movies, you would fill in #2.)

Adapted and reproduced by special permission of the Publisher, Psychological Assessment Resources, Inc., Odessa, FL 33556, from the Parenting Stress Index by Richard R. Abidin, Ed.D., Copyright 1990 by PAR, Inc. Further reproduction is prohibited without permission from PAR, Inc.
1. When my child wants something, my child usually keeps trying to get it.
2. My child is so active that it exhausts me.
3. My child appears disorganized and is easily distracted.
4. Compared to most, my child has more difficulty concentrating and paying attention.
5. My child will often stay occupied with a toy for more than 10 minutes.
6. My child wanders away much more than I expected.
7. My child is much more active than I expected.
8. My child squirms and kicks a great deal when being dressed or bathed.
9. My child can be easily distracted from wanting something.
10. My child rarely does things for me that make me feel good.
11. Most times I feel that my child likes me and wants to be close to me.
12. Sometimes I feel my child doesn't like me and doesn't want to be close to me.
13. My child smiles at me much less than I expected.
14. When I do things for my child I get the feeling that my efforts are not appreciated very much.
15. Which statement best describes your child?
   1. almost always likes to play with me,
   2. sometimes likes to play with me,
   3. usually doesn't like to play with me,
   4. almost never likes to play with me.
16. My child cries and fusses:
   1. much less than I had expected,
   2. less than I expected,
   3. about as much as I expected,
   4. much more than I expected,
   5. it seems almost constant.
17. My child seems to cry or fuss more often than most children.
18. When playing, my child doesn't often giggle or laugh.
19. My child generally wakes up in a bad mood.
20. I feel that my child is very moody and easily upset.
21. My child looks a little different than I expected and it bothers me at times.
22. In some areas my child seems to have forgotten past learnings and has gone back to doing things characteristic of younger children.

Adapted and reproduced by special permission of the Publisher, Psychological Assessment Resources, Inc., Odessa, FL 33556, from the Parenting Stress Index by Richard R. Abidin, Ed.D., Copyright 1990 by PAR, Inc.. Further reproduction is prohibited without permission from PAR, Inc.
23. My child doesn't seem to learn as quickly as most children.

24. My child doesn't seem to smile as much as most children.

25. My child does a few things which bother me a great deal.

26. My child is not able to do as much as I expected.

27. My child does not like to be cuddled or touched very much.

28. When my child came home from the hospital, I had doubtful feelings about my ability to handle being a parent.

29. Being a parent is harder than I thought it would be.

30. I feel capable and on top of things when I am caring for my child.

31. Compared to the average child, my child has a great deal of difficulty in getting used to changes in schedules or changes around the house.

32. My child reacts very strongly when something happens that my child doesn't like.

33. Leaving my child with a babysitter is usually a problem.

34. My child gets upset easily over the smallest thing.

35. My child easily notices and overreacts to loud sounds and bright lights.

36. My child's sleeping or eating schedule was much harder to establish than I expected.

37. My child usually avoids a new toy for a while before beginning to play with it.

38. It takes a long time and it is very hard for my child to get used to new things.

39. My child doesn't seem comfortable when meeting strangers.

40. When upset, my child is:
   1. easy to calm down,
   2. harder to calm down than I expected,
   4. very difficult to calm down,
   5. nothing I do helps to calm my child.

41. I have found that getting my child to do something or stop doing something is:
   1. much harder than I expected,
   2. somewhat harder than I expected,
   3. about as hard as I expected,
   4. somewhat easier than I expected,
   5. much easier than I expected.

Adapted and reproduced by special permission of the Publisher, Psychological Assessment Resources, Inc., Odessa, FL 33556, from the Parenting Stress Index by Richard R. Abidin, Ed.D., Copyright 1990 by PAR, Inc. Further reproduction is prohibited without permission from PAR, Inc.
42. Think carefully and count the number of things which your child does that bothers you. For example: dawdles, refuses to listen, overactive, cries, interrupts, fights, whines, etc. Please fill in the number which includes the number of things you counted.
   1. 1-3
   2. 4-5
   3. 6-7
   4. 8-9
   5. 10+

43. When my child cries it usually lasts:
   1. less than 2 minutes,
   2. 2-5 minutes,
   3. 5-10 minutes,
   4. 10-15 minutes,
   5. more than 15 minutes.

44. There are some things my child does that really bother me a lot.

45. My child has had more health problems than I expected.

46. As my child has grown older and become more independent, I find myself more worried that my child will get hurt or into trouble.

47. My child turned out to be more of a problem than I had expected.

48. My child seems to be much harder to care for than most.

49. My child is always hanging on me.

50. My child makes more demands on me than most children.

51. I can't make decisions without help.

52. I have had many more problems raising children than I expected.

53. I enjoy being a parent.

54. I feel that I am successful most of the time when I try to get my child to do or not do something.

55. Since I brought my last child home from the hospital, I find that I am not able to take care of this child as well as I thought I could. I need help.

56. I often have the feeling that I cannot handle things very well.

57. When I think about myself as a parent I believe:
   1. I can handle anything that happens,
   2. I can handle most things pretty well,
   3. sometimes I have doubts, but find that I handle most things without any problems,
   4. I have some doubts about being able to handle things,
   5. I don't think I handle things very well at all.

Adapted and reproduced by special permission of the Publisher, Psychological Assessment Resources, Inc., Odessa, FL 33556, from the Parenting Stress Index by Richard R. Abidin, Ed.D., Copyright 1990 by PAR, Inc.. Further reproduction is prohibited without permission from PAR, Inc.
58. I feel that I am:

1. a very good parent,
2. a better than average parent,
3. an average parent,
4. a person who has some trouble being a parent,
5. not very good at being a parent.

59. What were the highest levels in school or college you and the child's father/mother have completed?

Mother:

1. 1-8th grade
2. 9-12th grade
3. Vocational or some college
4. College graduate
5. Graduate or professional school

Father:

1. 1-8th grade
2. 9-12th grade
3. Vocational or some college
4. College graduate
5. Graduate or professional school

60. How easy is it for you to understand what your child wants or needs?

1. very easy,
2. easy,
3. somewhat difficult,
4. it is very hard,
5. I usually can't figure out what the problem is.

61. It takes a long time for parents to develop close, warm feelings for their children.

62. I expected to have closer and warmer feelings for my child than I do and this bothers me.

63. Sometimes my child does things that bother me just to be mean.

64. When I was young, I never felt comfortable holding or taking care of children.

65. My child knows I am his or her parent and wants me more than other people.

66. The number of children that I have now is too many.

67. Most of my life is spent doing things for my child.

68. I find myself giving up more of my life to meet my children's needs than I ever expected.

69. I feel trapped by my responsibilities as a parent.

70. I often feel that my child's needs control my life.

71. Since having this child I have been unable to do new and different things.

Adapted and reproduced by special permission of the Publisher, Psychological Assessment Resources, Inc., Odessa, FL 33556, from the Parenting Stress Index by Richard R. Abidin, Ed.D., Copyright 1990 by PAR, Inc. Further reproduction is prohibited without permission from PAR, Inc.
73. Since having a child I feel that I am almost never able to do things that I like to do.
74. It is hard to find a place in our home where I can go to be by myself.
75. When I think about the kind of parent I am, I often feel guilty or bad about myself.
76. I am unhappy with the last purchase of clothing I made for myself.
77. When my child misbehaves or fusses too much I feel responsible, as if I didn’t do something right.
78. I feel everytime my child does something wrong it is really my fault.
79. I often feel guilty about the way I feel towards my child.
80. There are quite a few things that bother me about my life.
81. I felt sadder and more depressed than I expected after leaving the hospital with my baby.
82. I wind up feeling guilty when I get angry at my child and this bothers me.
83. After my child had been home from the hospital for about a month, I noticed that I was feeling more sad and depressed than I had expected.
84. Since having my child, my spouse (male/female friend) has not given me as much help and support as I expected.
85. Having a child has caused more problems than I expected in my relationship with my spouse (male/female friend).
86. Since having a child my spouse (or male/female friend) and I don’t do as many things together.
87. Since having my child, my spouse (or male/female friend) and I don’t spend as much time together as a family as I had expected.
88. Since having my last child, I have had less interest in sex.
89. Having a child seems to have increased the number of problems we have with in-laws and relatives.
90. Having children has been much more expensive than I had expected.
91. I feel alone and without friends.
92. When I go to a party I usually expect not to enjoy myself.
93. I am not as interested in people as I used to be.
94. I often have the feeling that other people my own age don’t particularly like my company.
95. When I run into a problem taking care of my children I have a lot of people to whom I can talk to get help or advice.

Adapted and reproduced by special permission of the Publisher, Psychological Assessment Resources, Inc., Odessa, FL 33556, from the Parenting Stress Index by Richard R. Abidin, Ed.D., Copyright 1990 by PAR, Inc. Further reproduction is prohibited without permission from PAR, Inc.
96. Since having children I have a lot fewer chances to see my friends and to make new friends.

97. During the past six months I have been sicker than usual or have had more aches and pains than I normally do.

98. Physically, I feel good most of the time.

99. Having a child has caused changes in the way I sleep.

100. I don’t enjoy things as I used to.

101. Since I’ve had my child:
    1. I have been sick a great deal,
    2. I haven’t felt as good,
    3. I haven’t noticed any change in my health,
    4. I have been healthier.

STOP HERE — unless asked to do items below

During the last 12 months, have any of the following events occurred in your immediate family? Please check on the answer sheet any that have happened.

102. Divorce
103. Marital reconciliation
104. Marriage
105. Separation
106. Pregnancy
107. Other relative moved into household
108. Income increased substantially (20% or more)
109. Went deeply into debt
110. Moved to new location
111. Promotion at work
112. Income decreased substantially
113. Alcohol or drug problem
114. Death of close family friend
115. Began new job
116. Entered new school
117. Trouble with superiors at work
118. Trouble with teachers at school
119. Legal problems
120. Death of immediate family member

Adapted and reproduced by special permission of the Publisher, Psychological Assessment Resources, Inc., Odessa, FL 33556, from the Parenting Stress Index by Richard R. Abidin, Ed.D., Copyright 1990 by PAR, Inc. Further reproduction is prohibited without permission from PAR, Inc.
Appendix C:

Parenting Stress Index--Short Form
PARENTING STRESS INDEX
(Short Form)

Richard R. Abidin
University of Virginia

Directions:

In answering the following questions, please think about the child you are most concerned about.

The questions on the following pages ask you to mark an answer which best describes your feelings. While you may not find an answer which exactly states your feelings, please mark the answer which comes closest to describing how you feel.

YOUR FIRST REACTION TO EACH QUESTION SHOULD BE YOUR ANSWER.

Please mark the degree to which you agree or disagree with the following statements by circling the number which best matches how you feel. If you are not sure, please circle #3.

1 Strongly Agree 2 Agree 3 Not Sure 4 Disagree 5 Strongly Disagree

Example:
I enjoy going to the movies. (If you sometimes enjoy going to the movies, you would circle #2.)

1 2 3 4 5

Adapted and reproduced by special permission of the Publisher, Psychological Assessment Resources, Inc., Odessa, FL 33556, from the Parenting Stress Index by Richard R. Abidin, Ed.D., Copyright 1990 by PAR, Inc. Further reproduction is prohibited without permission from PAR, Inc.
1. I often have the feeling that I cannot handle things very well.
2. I find myself giving up more of my life to meet my children's needs than I ever expected.
3. I feel trapped by my responsibilities as a parent.
4. Since having this child I have been unable to do new and different things.
5. Since having a child I feel that I am almost never able to do things that I like to do.
6. I am unhappy with the last purchase of clothing I made for myself.
7. There are quite a few things that bother me about my life.
8. Having a child has caused more problems than I expected in my relationship with my spouse (male/female friend).
9. I feel alone and without friends.
10. When I go to a party I usually expect not to enjoy myself.
11. I am not as interested in people as I used to be.
12. I don't enjoy things as I used to.
13. My child rarely does things for me that make me feel good.
14. Most times I feel that my child does not like me and does not want to be close to me.
15. My child smiles at me much less than I expected.
16. When I do things for my child I get the feeling that my efforts are not appreciated very much.
17. When playing, my child doesn't often giggle or laugh.
18. My child doesn't seem to learn as quickly as most children.
19. My child doesn't seem to smile as much as most children.
20. My child is not able to do as much as I expected.
21. It takes a long time and it is very hard for my child to get used to new things.

Adapted and reproduced by special permission of the Publisher, Psychological Assessment Resources, Inc., Odessa, FL 33556, from the Parenting Stress Index by Richard R. Abidin, Ed.D., Copyright 1990 by PAR, Inc. Further reproduction is prohibited without permission from PAR, Inc.
22. I feel that I am: 1. not very good at being a parent,
2. a person who has some trouble being a parent,
3. an average parent,
4. a better than average parent,
5. a very good parent.

23. I expected to have closer and warmer feelings for my child than I do and this bothers me.

24. Sometimes my child does things that bother me just to be mean.

25. My child seems to cry or fuss more often than most children.

26. My child generally wakes up in a bad mood.

27. I feel that my child is very moody and easily upset.

28. My child does a few things which bother me a great deal.

29. My child reacts very strongly when something happens that my child doesn't like.

30. My child gets upset easily over the smallest thing.

31. My child's sleeping or eating schedule was much harder to establish than I expected.

32. I have found that getting my child to do something or stop doing something is:
   1. much harder than I expected,
   2. somewhat harder than I expected,
   3. about as hard as I expected,
   4. somewhat easier than I expected,
   5. much easier than I expected.

33. Think carefully and count the number of things which your child does that bother you. For example: dawdles, refuses to listen, overactive, cries, interrupts, fights, whines, etc. Please circle the number which includes the number of things you counted.

   1. 10+
   2. 8-9
   3. 6-7
   4. 4-5
   5. 1-3

34. There are some things my child does that really bother me a lot.

35. My child turned out to be more of a problem than I had expected.

36. My child makes more demands on me than most children.

Adapted and reproduced by special permission of the Publisher, Psychological Assessment Resources, Inc., Odessa, FL 33556, from the Parenting Stress Index by Richard R. Abidin, Ed.D., Copyright 1990 by PAR, Inc.. Further reproduction is prohibited without permission from PAR, Inc.