BUILDING LOCAL SKILLS: THE MULTITRAIT-MULTIMETHOD MATRIX IN PRACTICE

by

ANNA C. GERSH

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2014

MAJOR: EDUCATIONAL EVALUATION & RESEARCH

Approved by:

Advisor	**Date**

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

ACKNOWLEDGEMENTS

I am deeply grateful to my family. To my husband, Dave for his support and patience and to my son, Elvin for his love and understanding- it probably would have been easier if I was single and childless, but without you, there would be no one to share my joy in finishing. My friends, Paquetta Palmer, Eileen Storer-Smith, and Heidi Knabb have provided critical support on multiple fronts and I can't thank you enough. I am profoundly grateful to my advisor, Dr. Shlomo Sawilowsky, whose infinite patience, wit and deep understanding of evaluation methodology continues to be an inspiration. I would also like to thank Dr. Benjamin Kelcey, who has remained a consistent and highly valuable resource and guide to important methodological concepts, many of which I hope to eventually pursue in the context of this work. Dr. Monte Piliawsky came to this project when it was in its later stages, but has provided important critique regarding the larger policy context of this work. Dr. Gail Fahoome provided great inspiration and thoughtful critique in the early stages of this work. Her influence was felt throughout Wayne State's College of Education, but most importantly in the Educational Evaluation & Research Department. She will be deeply missed. Finally, I must express my deepest gratitude to Dr. Charles Smith, without whom this work would not be possible. He has provided an incredibly flexible working environment, supportive discussion, and an arena within which to make mistakes and develop my skills.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

The evaluation and measurement of constructs, or ideas that cannot be directly observed, is a persistent problem for the social sciences (Cronbach & Meehl, 1955; Crocker & Algina, 1986; S. Sawilowsky, 2007; Shadish, Cook, & Campbell, 2002). Constructs such as depression, happiness, and grit, are not directly observable but possess recognizable characteristics. Associated behaviors have been identified and classified, and scales have been developed to categorize their presence according to degree. In the absence of certainty, triangulation among both qualitative and quantitative methods of measurement is used to build an argument for or against the presence of such notions (Erzberger & Prein, 1997; Mathison, 1988; Campbell & Fiske, 1959). Constructs are used to explain reality, including individuals and places (Sawilowsky, 2007). As this concept relates to cities, for example, they may be categorized or designated as safe, accessible, or poor, meaning these anthropomorphic attributes convey ideas of safety, accessibility, or poverty.

One area that has been especially dependent on definition by construct is education. Schools are also anthropomorphized as being good, bad, and safe. Opportunity is offered, or choices are limited. An ostensibly simple list of desirable characteristics for one's public school system might include multiple constructs (e. g., safe, close to home, good teachers, strong leadership, effective partnerships, supportive community, high academic standards) ultimately, the central question in the minds of those choosing a school is, "Is this a good school?" More specifically, can the seeker be certain of an adequate level of quality. Although there are multiple

accreditation bodies to offer assurances for some types of constructs (e.g. teacher quality or academic standing) , many formal and informal assessments exist to offer additional support to families and other stakeholders interested in determining school quality (greatschools.org, 2013; OECD, 2013). The concept of quality is a construct which itself subsumes multiple constructs. However, in the context of education or the generalized social-behavioral outcomes associated with varying educational enterprises, quality might be further explicated as a set of environmental conditions which lead to effectiveness. In spite of this great variety of resources for unraveling the meaning behind these constructs - most notably that of quality, researchers, governments, funding organizations, and administrators in many educational enterprises continue to seek information about quality in learning environments, and they are concerned with how quality can be further developed.

In January 2013, the Bill and Melinda Gates Foundation published the final report from the three year Measures of Effective Teaching (MET) study, which sought to identify quantifiable aspects of effective teaching which could be linked to improved academic outcomes for students and then used to support targeted professional development. Among their recommendations were multiple measures of teacher practice, including student surveys, student achievement data, and classroom observations. Multiple measures were found to be more reliable and less volatile over time than student achievement data alone (Cantrel & Kane, 2013). Additionally, some researchers are beginning to devise systems to support and incentivize quality improvement via performance management and public sector accountability (Camm & Stetcher, 2010). Such Performance-Based Accountability Systems (PBAS) seek to

improve quality by setting clear standards and providing rewards for meeting those standards. United Way of Kansas City and Prime Time, in Palm Beach County, Florida are two examples of after school providers who are employing such systems. Both networks link performance incentives with fidelity to a quality improvement model. Implied by both these approaches is an increase in data management on practitioners.

Given the variety and nature of applicable constructs to describe school quality, and the near universal use of many educational enterprises (as of 2011, the US national average for net enrollment rate (NER) for primary education was 97% and for US secondary education the NER was 91% (UNESCO, 2011), it is no surprise that most communities' investment in educational enterprises (e.g. public schools, affiliated academic and social support services, after school programs, community-based education) is substantial. At the federal level, 1.3 billion was allocated in 2014 alone for preschool education (Statistics, 2012b), with 591.3 billion projected expenditures for 2013-2014 (Statistics, 2012a). The State of Michigan, alone projected a total of just over 15 billion for the school aid fund for fiscal years 2014-2015 (Snyder & Nixon, 2014). Given these substantial investments, it is not unreasonable that some assurances of effectiveness have been sought. Legislation including; No Child Left Behind (NCLB), Race To The Top (RTTT), Common Core, and programs like the National Baccalaureate Program are examples of attempts to promote effectiveness through accountability for school-day programs (School-day programs refer to those programs directly connected to public, educational experiences that take place during the school-day). However, criticisms particularly of NCLB (by far the largest comprehensive attempt to promote effectiveness among American educational institutions), suggest

decision makers in traditional educational institutions continue to struggle with the problem of such assurances (Fusarelli, 2004; Hall, 2013).

Educational enterprises share common characteristics. They seek to develop skills in the participants that they hope will transfer to other settings. They struggle with limited funding, often dependent upon soft money, such as foundation or government grants; and the potential success of their efforts is in part dependent upon a qualified and diversely skilled staff (Heyneman & Loxley, 1983; Phillipsen, Burchinal, Howes, & Cryer, 1997). Skill development of participants and the development of staff capacity are essential features of what is generally understood to mean quality programs among educational enterprises. From this point on, the term quality will be used as a construct to subsume the quantifiably measurable attributes of the term, effective. For example, where effective might mean a state education agency's ideal graduation rate or a certain percentage of students meeting a selection of preset college readiness indicators, the implication of a quality program is that having met the standards of quality as defined by the elements discussed in this paper, educational enterprises will be more likely to meet the more quantifiable standards associated with effectiveness..

The construct of quality is served by a dizzying variety of professional development options and Quality Improvement Systems (QIS) (QIS refers to large scale program interventions intended to address multiple aspects of program quality, typically conducted over multiple years). The Professional Development Sourcebook compiled by Education Week includes 42 subject categories offered by 740 organizations (Education Week, 2014). In terms of large scale QISs, or coordinated efforts (often among multiple service providers) to promote quality improvement, there are several

notable examples. Perhaps the best known of these is the Prime Time Initiative in Palm Beach County, FL (Spielberger, et.al, 2009) a system designed to promote quality and availability of after school programs, started in 2005 (The David P. Weikart Center, developers of the Youth Program Quality Intervention (YPQI), is an ongoing participant in this QIS, as an intermediary service provider. Data from the YPQI validation study (Smith et al, 2012) is evaluated in this paper. The Youth Program Quality Assessment (Youth PQA) is a key aspect of the YPQI and an adapted version of the Youth PQA is a central feature of the data collection in the Palm Beach QIS).

As the industry around quality improvement has grown, so too have the expectations for educational enterprises, especially public schools. Expectations of continuous quality improvement and optimum performance have been set for educational enterprises and high stakes decisions are made based on educational enterprises' ability to demonstrate measurable growth in quality. As educational budgets continue to shrink and programs once thought to be effective must demonstrate improvement in a range of quality measures or be closed, there has been no shortage of interventions, often commercial products generated by for-profit entities, which claim to affect the quality of those programs.

Given the expectations of stakeholders, and the variety of professional development and other accountability efforts, it behooves decision-makers in educational enterprises to be better equipped to evaluate the evaluators and potentially even make their own measures. Evaluating quality in such organizations has frequently been as much a matter of art as of science, although science has provided some useful ways of supporting such determinations.

With respect to discussions about quality programming in a learning setting, the literature provides a rich, if slightly overwhelming, palate of suggestions. Evidence suggests quality is improved with increased focus on the development of social-emotional skills for youth (Durlak & Weissberg, 2007; Jones & Bouffard, 2012); improving teacher quality (Darling-Hammond, 2000; Rockoff, 2004); expanding individualized instruction and tutoring opportunities (Cook, et al, 2014) and "what goes on in the classroom and the overall culture and atmosphere of the school" (Mayer, 2000). Additionally, studies such as the MET are providing empirical support for the increased use of multiple measures to identify quality in teaching, widely recognized as one of the critical factors in improving educational outcomes for children and youth. Public discussions about the poor quality of urban schools reveal nothing of the complexity involved in determining the nature of quality. Meanwhile public expectations for evidence of quality are increasing.

Multitrait-Mulitmethod Matrix

What we know from the MET study and other studies that identify desirable characteristics we want to reproduce in educational enterprises is that quality can be indicated in a variety of ways and multiple methods of measuring quality provide a more reliable evaluation of that construct than any single approach. In order to unravel the nature of a construct like quality with the goal of improving the ability of frontline decision-makers (e.g. principals, administrators, district level researchers, local evaluators, etc.) to choose from a wide variety of quality improvement methods, it is essential to identify a method that is sophisticated enough to be embraced by methodologists, yet simple enough to be conducted and interpreted by a graduate

student. A powerful but straight forward method of evaluating constructs was devised by Campbell and Fiske (1959). According to Google Scholar, as of July, 2014 their article was cited 13,016 times and is considered to be among the most influential papers published in the field of psychology (Sternberg, 1992).

Cambell and Fiske (1959) argued that given the influence of method factors (variability in scores due to characteristics of the measurement method) it is critical to analyze constructs, and the traits associated with constructs, using a variety of methods. Comparisons among the different method and trait correlations reveal the strength of a given construct in terms of both convergent and discriminant validity. If different methods measuring the same trait are highly correlated, this is evidence for convergent validity, because regardless of the method used to measure the trait, its existence is evident. If different traits measured by either the same or different methods are not highly correlated it is evidence for discriminant validity, because traits distinguish themselves regardless of the method used to measure them. Analysis of these correlations reveal the strength of the overarching construct and taken together provide compelling evidence for construct validity.

The MTMM is in some respects a simple tool. It takes the most fundamental concepts of measurement, those of reliability and validity, and uses as its central formula the Pearson Product Moment Correlation (r) (Pearson, 1895) to methodically build an argument for the existence of the construct of interest. These concepts are some of the first aspects of measurement theory learned by every graduate student (and not a few undergrads) in just about every area of higher education. These three concepts, reliability, validity and r (so familiar and ubiquitous it is known simply as the

correlation) are so rudimentary they are described by the third chapter in just about every social science textbook known, yet the MTMM remains a challenging procedure to analyze (Kenny & Kashy, 1992; Maas, Gerty, Lensvelt-Mulders, & Hox, 2009).

Reliability and Validity

Interpretation of the MTMM is complicated by the illusiveness of the foundational concepts of reliability and validity. The first set of correlations employed in the MTMM is the monotrait-monomethod correlations. These measurements are reliability measurements of one trait, measured by one method. An acceptable measure of reliability is widely known to be a necessary, but insufficient characteristic of validity (Nunnally, 1978). The reliability of a test was once understood to be the measure's ability to reproduce similar scores, by successive administrations of a measure (test-retest reliability); by demonstrating item equivalence (parallel forms reliability); or by demonstrating internal consistency (that the items are consistently measuring the same thing, usually evaluated by split-halves reliability, or coefficient alpha).

In recent years, reliability has come to be associated with the consistency of the scores rather than the test (AERA, 1999). On the basis of a literature review Sawilowsky (2000) rejected that approach, and noted there is still some question even among score-reliability advocates about the proper application of the concept. Ultimately, the concern of both vertical and lateral consistency remains the fundamental focus of reliability. Cronbach's alpha (also known as the Standardized Item Alpha) is the most frequently cited statistic as evidence of score reliability, (c.f. J. Smith, 2011), however it measures only one type of reliability. Cronbach's alpha measures internal consistency

by mimicking a split-halves reliability whereby each item is correlated with other items in the scale and the average correlation results in the alpha statistic. Because the alpha measures only internal consistency, high correlations are thought by many to represent a lower bound of the reliability (Cortina, 1993; Cronbach & Shavelson, 2004; Sijtsma, 2009; McCrae, Kurtz, Yamagata, & Terracciano, 2011). Some evidence suggests that alpha may also over estimate reliability in some circumstances (Raykov, 1998) and other statistics may provide more accurate estimates of the true reliability (Bliese, 2000) (Sijtsma, 2009; McCrae et al., 2011).

Validity presents even greater interpretative challenges, especially with respect to current trends in measurement and most especially in the field of education. It is inappropriate, from a classical measurement perspective, to call a test of any type valid. However, it is appropriate to ascribe validity to the usage of the test. In fact, the Latin origin of the word, *validere*, means strong. In litigation, validity is associated with legal arguments, and is tantamount to establishing truth. However, the meaning of valid in a scientific measurement context is rarely that certain and its colloquial use disguises its dependency on multiple sources of highly vulnerable and variable information to build its case. In the sciences, validation is a process consisting of multiple steps intended to build a body of evidence supporting use. Validity is “the truth of, correctness of, or degree of support for an inference” (Shaddish, Cook, & Campbell, 2002). Validation is therefore context specific. This definition might be applied to a measurement tool, a theory, or a process. In each case, the process of establishing validity is necessarily tailored to the research topic.

For most of the 20th century, determining validity in the social sciences was organized into three categories of argument; criterion-based (how well a test correlates with other tests meant to measure the same thing), content-based (expert evaluations of the content of the test) and construct-based (does the test measure what it is supposed to measure). As time has gone on, these categories have been further subdivided to account for varying uses of the test (predictive vs. concurrent validity); consequences of the results of testing (consequential validity); and potential authoritative acceptance by non-expert users or examinees (face validity) (Messick, 1989; Sireci, 2009; Cizek, Bowen, & Church, 2010). However, both consequential and face validity have been called into question by researchers as subjective and unquantifiable (Popham, 1997; Mehrens, 1997; S. Sawilowsky, Personal Communication, 2012)

Convergent validity is used to describe collections of evidence that, when considered together demonstrate validity for a given purpose. Discriminant validity is used to describe evidence with low correlations that demonstrate an item or trait or scale's uniqueness. These interpretations of validity account for a variety of both scientific and non-scientific justifications for use. The most recent iteration of the American Psychological Association's Standards for Educational and Psychological Testing (AERA, 1999) identifies construct validity as the essential determination of validity in that establishment of construct validity can subsume all previous definitions of validity. Among all previously defined categories of validity, construct validity arguments are built with the most scientifically (quantitatively) defensible arguments (AERA, 1999). This definition of legitimate validity also has its detractors, and authors continue to challenge its limitations (Messick, 1989; Cizek et al., 2010; Lissitz & Samuelson, 2007).

The MTMM focuses on two types of validity, convergent and discriminant validity and these are determined via correlations between scores. High convergent validity among methods measuring a single trait of the construct of interest and high discriminant validity between measurements of different traits using both similar and different methods combine to provide evidence of construct validity.

The availability of quantitative data to fuel system change has brought research level conversations into the boardroom. Once the near-exclusive realm of the psychometrician, discussions of measurement selection based on the reliability and validity of measures are being conducted among a variety of minimally qualified decision-makers, making the ability of systems-change agents to describe the true value of their measures in scientific terms and with a nuanced understanding of the essential terms all the more prescient. At the same time, this presents an opportunity for instructional leaders to contribute practical experience and understanding of the instructional environment. Recent changes in in funding opportunities like Race to the Top have emphasized the importance of principals and other educational leaders as managers of large bodies of data for which they may be held accountable. Heads of educational enterprises are being offered an opportunity to deepen their involvement with instructional leadership through highly developed quality improvement systems and translate their knowledge of professional skills into quality constructs built from observable data. This more highly developed role of instructional leader and data manager requires a strong working knowledge of basic research concepts, in so far as the ability to produce evidence of effectiveness is a critical feature of fund development which also fuels effective capacity development. Quality improvement systems,

professional development programs, and other products designed to help educational enterprises meet community, state, and federal expectations are available to instructional leaders and others charged with managing and sustaining educational enterprises. Instructional leaders are expected to support fund development. They do this by writing grants, galvanizing community support, and more and more by the collection and management of large amounts of data. Certainly many larger educational enterprises are supported by specially skilled evaluation and research support teams, but smaller organizations may not have this same support network, yet expectations for data driven accountability practices are the same for all educational enterprises. It is now imperative for administrators of all educational enterprises to expand their research and evaluation skills such that organizational size will not impede access to resources.

In spite of the fluid nature of the basic terms, the MTMM offers an accessible framework for beginning to examine the relationships among traits as they concatenate to form constructs. Campbell and Fiske's (1959) decision rules provide a way of evaluating relationships among the traits and methods that can begin to be assessed with a basic working knowledge of research. The difficulty is in the extraction of a final determination of whether or not the traits, in combination, evidence construct validity to the satisfaction of scientific research standards. One approach to the analysis of the MTMM is confirmatory factor analysis (Kenny & Kashy, 1992; Maas et al., 2009). Multilevel modeling has also been shown to be a useful method in certain circumstances (Maas et al., 2009). These techniques are beyond the scope of the typical consumer of educational improvement measures. The I test (Sawilowsky, 2002), however, can be implemented without the highly specialized technical knowledge

necessary for such procedures, and being distribution free, it does not have the required underlying assumptions as does confirmatory factor analysis or multi-level modeling.

The Youth Program Quality Intervention (YPQI)

The YPQI is a quality improvement system currently in use in over 95 state, county and city expanded learning networks, accounting for well over 3,500 individual sites. Included among some of the more well-known participating networks are Boys and Girls Clubs of America, United Way, and 21st Century Community Learning Centers. The YPQI is a product of High Scope Educational Research Foundation (Smith & Hohmann, 2005; CYPQ, 2012) and the David P. Weikart Center for Youth Program Quality (Smith et al, 2012), the legacy of these organizations includes the Perry Preschool Project (The Perry Preschool study has provided over three decades of longitudinal evidence for the benefits of quality preschool on lifetime earnings, health, and delinquency and crime among other factors (Schweinhart & Weikart, 1997). The intervention is based in the Active-Participatory method, first developed at High Scope, and later refined as a measure of system quality at the point of service by researchers at the Weikart Center. The YPQI is a system intervention that includes both internal and external evaluation using the Youth PQA observational measure (The Youth PQA is designed for students in 4th-12th grades. A School-Age PQA is also available, designed for instructional situations catering to students in K-6th grades); internal and external coaching using a proprietary method (Quality Coaching/Observation-Reflection); proprietary training (Youth Work Methods); and improvement planning, which focuses on use of the data collected on the observational measure for system improvement.

Although it is currently making inroads into the regular school day, YPQI was designed with after school programming in mind. After school programming is distinguished by several features that make it different than traditional school-day programming. Notable among these is that after school programming participation is typically voluntary (Cross, Gottfredson, Wilson, Rorie, & Connell, 2010) and may disproportionately draw at-risk students. As such, its generalizability may be limited with respect to regular school-day programming.

The YPQI validation study is described in Continuous quality improvement in afterschool settings: Impact findings from the Youth Program Quality Intervention study (C. Smith, Akiva, T., Sugar, S., Devaney, T., Lo, Y., Frank, K., Peck, S., Cortina, K., 2012). Funded by the William T. Grant Foundation and using data collected between 2006 and 2009, the study attempted to evaluate impact and implementation issues associated with the YPQI, as well as several field level questions related to policy. In the intervention group, the study was able to demonstrate an impact on manager focus, specifically that managers were found to be 7.29 times more likely to focus on instructional issues, following the intervention. It was also determined that staff tenure increased in the treatment group, following the intervention, and; perhaps most significantly, staff engagement in continuous improvement practices increased following the implementation year and also after the follow-up year (Smith, et. al., 2012).

Purpose of the study

The MTMM matrix will be used in a secondary data analysis on two sources of data from the third year of the Youth Program Quality Intervention (YPQI) impact study

(Smith et al, 2012). The matrix will incorporate a survey of staff and data from an observational measure (the Youth Program Quality Assessment (Youth PQA) to evaluate the construct of quality as it has been defined by previous validation work on the YPQI. Quality, as it has been defined in the YPQI, in terms of the point of service (POS- Point of service is the point where adults deliver instruction to youth during program hours (Smith et al., 2012)), is made up of four traits: Safe Environment; Supportive Environment; Interaction; and Engagement. These traits have been included in both the observational measure and the youth survey as separate scales. In the staff survey these ideas have been used to inform the items. The *I* test (Sawilowsky, 2002) will be used to assess construct validity by evaluating the presence of trend in the data. Following the analysis, an examination of the usefulness of the proposed procedure and thoughts on its value will be discussed.

Importance of the study

As a result of this study, the possibility of adapting existing research methodology, previously the exclusive domain of methodologists, to support the professional development of site level administrators will be evaluated. The expectations of site-based administrators and school district researchers to collect, manage, and incorporate student data with the intention of using this data for accountability or improvement efforts is growing. So too is the industry around data driven accountability and improvement such that site-based administrators need to develop capabilities, first as discerning consumers. Site-based administrators, perhaps in cooperation with local evaluators, may also be able to use this method to identify existing constructs in the unique practices associated with their site, staff or other local

procedures and thereby contribute to the larger research conversation by the preliminary validation of local practices. This study is a small step toward the development of a set of evaluation tools that can be employed without the necessity of highly specialized methodological training.

The data sources used in this study, specifically teacher observational data and staff surveys are the same data sources identified by the MET study as necessary to the accurate evaluation of effective teaching as well as the same data sources identified by recently enacted federal grant programs (Race to the Top) as critical to determining the effectiveness of educational enterprises. Both the MET and RTTT also include youth surveys to support evaluation studies, as does the YPQI study. Youth surveys were originally meant to be a part of the analyses, however the differences between the W1 and W3 data with respect to item phrasing and measurement scale were dramatic enough that the Day of Youth Survey had to be omitted as the third measurement method. Item comparisons for the W1 and W3 Day of Youth Surveys have been included in Appendix B (see Tables B5-B7). This study illustrates a method that might be used to evaluate these multiple data sources, using both intuitive logic and fundamental statistical methodology.

The goal of developing such skills is not meant to suggest that site-based administrators and school district researchers might replace traditional or external research professionals, but rather that given the expanding use of data in terms of accountability and improvement efforts (especially in education) that the development of these skills for site-based decision-makers may benefit both local service providers and research professionals. For local service providers, the benefits might be in terms of

increasing access to resources by developing administrators' fund development and measure selection, and measure development skills. For research professionals, specifically university level researchers who often work in partnership with local service providers, the benefits might be in terms of improved data management at the site level.

Establishing construct validity using the MTMM benefits these site level research professionals by developing sensitivity around measure development and enabling practitioners to strengthen an argument for a particular or preferred site-specific practice based on a deeper understanding of established research methodology. It increases the potential for practice to inform research by providing a simpler and robust method of thinking about construct validity and may support the development of evidence-based, site-based performance measures by expanding practitioner interest in and accessibility to research methodology.

Finally, developing local skills around validity investigations will support appropriate measure selection and the ability to compare and combine multiple measures of effectiveness as well as supporting deeper investigations of practice at the local level.

This paper also provides an opportunity to use the rank-based *I* test in an applied educational setting. This test supports the development and analysis of constructs by local education professionals by providing an accessible method of examining measurement and practice methods. It supports local evaluators by providing a tool to begin conversations about educational constructs as they are realized through measurement methods.

Limitations

This *I* test approach provides evidence for construct validity, but it is not a complete picture of validity, or of the usefulness of a given measure. Instead, this procedure delivers support for further investigation; for the methodologist, evidence for more sophisticated statistical modeling using confirmatory factor analysis (CFA) or multilevel modeling; for the program evaluator or administrator, evidence for continued interest in the measure and additional qualitative and quantitative evidence of the purpose and usefulness of the measure.

The MTMM is a similarly restricted measure. Although the matrix was originally constructed to accommodate any type of trait measure (G. Smith, 2005), modern statistical methods might identify problems associated with the potential interpretation of traits at multiple levels of analysis, also called trait, or psychometric isomorphism (Tay et al., 2013). While such limitations are ultimately important considerations in the confirmatory stages of measure development and analysis, this paper is concerned with the expansion of research tools at the practice level. To that end, the procedures outlined in this paper are intended to support initial stages of analysis and development at the local level with the understanding that research and development is inherently iterative and the development of skills at all levels supports the ongoing refinement of tools and practices throughout the educational measurement field.

Chapter 2

Literature Review

Citing *Psychological Bulletin's* "Top 10 Hit Parade", Sternberg (1992) detailed the impact of Campbell & Fiske's (1959) article on the Multitrait Multimethod Matrix. At the time, this was the most frequently cited article in Psychological Bulletin's history, with over 2000 citations. As of July, 2014, Google Scholar identified 13,016 citations and that number continues to grow.

Fiske & Campbell (1992) addressed the sustained popularity of the MTMM. In a short response paper entitled, "Citations Do Not Solve Problems" the authors identified reasons they believed, in spite of vastly improved computer power enabling developments in analytic techniques, why the MTMM continued (and continues) to be used by researchers to evaluate and provide evidence for construct validity and as a way to parse method effects. First they suggest that the MTMM is easy. "It combined obvious desiderata with an explicit how-to-do-it recipe…" (p.393) and further that "The validational recipe did not require that any measure be treated as a perfect criterion, thus meeting the needs of the great majority of personality trait or attitude measurers" (Fiske & Campbell, 1992, p.393). These benefits certainly support wide appeal for the procedure. The first point suggests that the matrix procedure might make aspects of research more widely accessible. The second suggests a reasonable and welcome simplification of the validational process, especially in the case of educational and behavioral researchers.

Fiske (1982) points out that the MTMM addresses an essential need in the social sciences. It provides a framework within which one may unravel the overlapping nature

of abstract concepts so prevalent in the study of human behavior. By examining correlations among methods and traits of constructs one might more clearly distinguish between ideas such as determination and happiness. The MTMM suggests that careful selection of traits combined with appropriately discriminating methods of measuring these traits can support construct validity by demonstrating convergence across methods and can also help identify the variance associated with method through discrimination across traits (Fiske, 1982).

Although its benefits suggest simplicity, it continues to be plagued by difficulties. Perhaps the most notable is that after 55 years, there is still no ideal way to evaluate the matrix to produce a single determination of validity. Campbell & Fiske's decision rules provide a foundation from which one might be able to determine support or lack of support for either validation or obvious interference of method factors, but in terms of a definitive answer, the literature is silent. Evaluation approaches are burdened by method-trait interactions (Campbell & O'Connell, 1982; Campbell & O'Connell, 1967; Bagozzi, Yi, & Phillips, 1991; Podsakoff, MacKenzie, & Podsakoff, 2012); out of range estimates and convergence problems (Putka et al., 2011; Lance, Woehr, & Meade, 2007; Kenny & Kashy, 1992) ; software limitations (Maas et al., 2009); and problematic data (Cote, 1995).

Method effects, or the variance associated with measurement method is the reason for the MTMM. Bias associated with measurement method is a well-known problem in research and the various threats associated with method bias have been addressed extensively in the literature (Campbell & Fiske, 1959; Campbell & Stanley, 1963) ; (Shadish et al., 2002); (Podsakoff et al., 2011). The matrix works by

triangulation, specifically when different methods find evidence of the same trait, one can say that trait or construct is evident. However, it is known that measurement methods possess variability, unique to the methods by which they are measured, as in the case of item wording or other instrument effects that influence the outcome of an experimental trial. Campbell and O'Connell (1967) pointed out that when traits are measured by the same method, the correlation between traits is positively influenced by the shared measurement method. They called this differential augmentation; "The higher the basic relationship between the two traits, the more that relationship is increased when the same method is shared" (Campbell & O'Connell, 1982, p.95), in other words, if the correlation starts out high, shared method will increase that correlation. They also found that traits measured by different methods might also result in an attenuated, or weakened correlation, again influenced by the interaction of method with trait.

Confirmatory Factor Analysis has been the preferred method of evaluating the MTMM. However it often results in model misspecification and the problems are compounded when the number of constructs is high (Woehr, Putka, & Bowler, 2011). In a Monte Carlo study conducted by Lance et al. (Lance, Woehr, & Meade, 2007), 500 sample matrices were analyzed each using three different CFA models. In this way, Lance et al. were able to see which of those 1500 matrices were able to produce convergent and acceptable solutions. They found that whether or not the data conformed to the original fit specifications of the model, "CFAs based on multiple dimension factors converged with an admissible solution for only 57% of the data matrices" (Woehr et al., 2011). They also found that among the acceptably converged

models, several typical indicators of appropriate model fit suggested good fit even when the fitted model did not match the population data on which the model was based. Further, as the chi square is one of the main fit indices used to evaluate the CFA, the researcher must also be concerned with the number of traits and methods used in the model. The chi square is sensitive to sample size and may render a significant result simply due to the researcher's overfitting of the model (Bagozzi & Yi, 1990). Kenny & Kashy (1992) also point out that, as the number of latent constructs increases, so does the potential for unstable solutions.

Maas et al. (2009) also examined both straightforward (9x9 trait-method matrix) and complex MTMM data (Big 5 personality data) using CFA and multilevel modeling. While they determined CFA was useful for modeling straightforward data (specifically one measure per trait-method unit), they found that when the data was more complicated (e.g. multiple, interchangeable raters) multilevel modeling was more flexible. Although they do not directly address the problems of model misspecification, they point out that SEM software, as of 2009, had not yet caught up to the challenges of complex data sets which are typical of education data. They also showed that "the multilevel approach can be viewed as a confirmatory factor model with additional restrictions on the factor loadings" (Maas et al., 2009, p.76) but that the unrestricted model was preferred. The different approaches serve different circumstances, but neither is without notable limitations.

The Direct Product Model (Browne, 1984) attempts to address the multiplicative effects of traits and methods interactions identified by Campbell and O'Connell (1967; 1982) by employing covariance matrices associated with the methods and traits and

rendering a disattenuated covariance matrix. Browne (1984) developed proprietary software to estimate fit parameters including standard errors and a chi square goodness of fit index. However, reliance on the chi square suggests that sample size sensitivity associated with the chi square may still be a problem. Additionally, the DPM is still dependent on the same sorts of comparative decision rules used by Campbell & Fiske and “ambiguity arises when one must decide how much variance is sufficient for attaining convergent validity” (Bagozzi & Yi, 1990, p. 433). So while more directly specified than the Campbell and Fiske approach, it would seem that estimation is still largely heuristic, and the evaluation procedure is decidedly more complicated.

More recently, Woehr et al. (2012) suggested the use of Generalizability Theory (L. Cronbach, Rajaratnam, & Gleser, 1963) to evaluate the matrix. Woehr’s procedure includes use of either a univariate or a multivariate model, supported by structural equations modeling. Because G-Theory models are ANOVA based, they are highly constrained compared with CFA. The covariance matrices are based on average covariances, taken across trait-method units. These produce common fit indices across the trait-method units, making model fit difficult and often inhibiting convergence. Additionally, ANOVA-based models are subject to the standard parametric assumptions of independence, linearity and homoscedasticity, often difficult to meet for data associated with educational enterprises.

Ultimately the researcher is forced to either embrace a considerably less than ideal evaluation of the matrix or disregard its use altogether. This paper suggests that perhaps the best use of the matrix is in the hands of practitioners and local evaluators rather than methodologists. The persistent popularity of the MTMM may be due at least

in part to its simplicity, as Campbell and Fiske (1992) suggest but it is also certainly due to its intuitive logic. Evaluation and quality improvement in educational enterprises is more and more the purview of network leaders and principals often, but not always, working closely with district level evaluation teams. These practitioners possess authentic experience evaluating abstract constructs such as quality via practical daily exposure. Further, evaluation and quality improvement are here to stay. Decision-makers will not abandon their desire to improve expected outcomes via quality improvement efforts any time soon and the industry that continues to develop around improvement will also continue to develop new research-oriented products designed to serve practical purposes. It can only better serve educational enterprises to have more local decision-makers involved in the selection, and potential creation, of evaluation tools. The MTMM provides an opportunity for practitioners to bridge the gap between research and practice. What may be “eyeballing” (Campbell & Fiske, 1992, p. 394) in the mind of a methodologist, may be a “sound first step” in the hands of a practitioner, “especially when one has an extended research program and sees the particular matrix as only a step toward constructing an improved set of measuring procedures. Presumably, one is not going to stop with the particular matrix but is going on to further study of the variables by methods improved by carefully interpreting the results at hand” (Campbell & Fiske, 1992, p. 394).

As thoughtful evaluation of data by a practitioner and daily participant in both the creation and collection of the data may serve the higher level researcher or methodologist by supporting thoughtful and complete data collection at the site level, experience with the decision rules strengthens understanding of the construct and the

data for the practitioner. Cote (1995) suggested that data quality, including management of outliers, may be the single most common cause of estimation problems when evaluating MTMM data via CFA. It may be that including practitioners more deeply in data management might help alleviate some of the problems associated with data quality and help facilitate analyses at higher levels. Practitioners are better able to see data in terms of the larger context. It becomes truly meaningful professional development in that extending local decision-makers' knowledge of the constructs used to evaluate educational programs empowers local decision makers, strengthening investment in meaningful evaluation by making the evaluation process transparent and part of the practical work. Connecting professional development to daily work has been identified by the Eisenhower Professional Development Program (Garet, Porter, Desimone, Birman, & Kwang, 2001) as one of four critical ingredients of effective professional development. Expanding evaluation and methodological skills for local decision makers in a period of increasing reliance on evidence-based outcomes may prove useful to educational enterprises at many levels.

Practitioners may provide useful and economical gate-keeping in the service of quality improvement by facilitating thorough data collection. Using the MTMM, it may be possible for practitioners to provide preliminary analysis of evaluation data in terms of construct validity, and once preliminary analysis is complete, the I test (Sawilowsky, 2002) provides an opportunity to further support analysis at the local level by means of a quick distribution-free test for trend that contributes evidence of construct validity.

The I test is a rank-based statistic that averages the values in the MTMM to find the minimum, median and maximum values within the reliability diagonal; the validity

diagonal; the heterotrait-monomethod triangle; and the heterotrait-heteromethod triangle (See Table 1).

Table 1

Matrix

Staff				YPQA		
	Staff Supp	Staff Int	Staff Eng	YPQA Supp	YPQA Int	YPQA Eng
Staff S	(.51)					
Staff I	**.185****	(.57)				
Staff E	**.166***	**.516****	(.74)			
YPQA S	*.020*	.088	.171*	(.78)		
YPQA I	.028	*.184***	.177*	**.552****	(.80)	
YPQA Eng	.052	.-.004	*.131**	**.476****	**.494****	(.75)

*Note: * indicates significance at the .05 level. ** indicates significance at the .01 level. In the matrix, the reliability diagonal, also identified as the monotrait-monomethod values, are the reliabilities of each of the trait-method units. In the matrix they are identified by parentheses. The validity diagonal consists of the monotrait-heteromethod values. These are the correlations within traits and across methods. High correlations on the validity diagonals are indicative of convergent validity. These are identified by italics. The heterotrait-monomethod values are the correlations between traits and within a single method. These are printed in bold. The heterotrait-heteromethod values are those correlations between different traits and different methods. Comparatively lower correlations among the heterotrait-heteromethod values suggest discriminant validity. These values are shaded in gray.*

Once these values are determined, they are evaluated by a set of decision rules based on simple comparative logic. Ideally, the values in the reliability diagonal will be higher than those in the validity diagonal, which will be higher than those in the heterotrait-monomethod triangle, which will be higher than those in the heterotrait-heteromethod triangle. At each of these levels the coefficients will have been reduced to

minimum, median, and maximum values. The null hypothesis states that the values at each level have a random order H0: f(x1)=f(x2)=…=f(xn). The alternative hypothesis demonstrates ordered trend H1: f(x1)<f(x2)<…f(xn), which is the best case scenario, demonstrating the strongest evidence for construct validity. In cases where trend is less evident, significance is determined by the number of inversions, or reversals in order of the data. At nominal α =0.05, for a 3x3 matrix, data can demonstrate up to 14 inversions and still be considered significant evidence of construct validity (Sawilowsky, 2002). As is the case with rank-based statistics, some sensitivity is sacrificed in favor of clarity, however, for the purposes of this procedure, which is meant to support local evaluators and others, not specifically trained as methodologists; this procedure represents a significant improvement in local decision makers' evaluation skills and may be used to support decision making at the local level.

In terms of Type I error rate and statistical power, the *I* test performs satisfactorily when applied to matrix layouts similar to those described in Campbell and Fiske (1959). Sawilowsky replicated the results of Campbell and Fiske's data sets (1959, Table 12, p. 96 and Table 2, p. 86) using 3x5 and 2x4 matrix layouts, respectively.

Cuzzocrea (Cuzzocrea, 2007) found that the I test demonstrates increasing Type I error rates with increasing data points. This was predicted by Sawilowsky (2002), because the *I* test violates the independence assumption. The minimum, median and maximum scores are collapsed versions of the full data set. The resulting internal correlation structure of the scores is revealed by the increasingly conservative Type I error rate as additional data points are added. As might be expected, the relative efficiency was found to decrease with increasing data points.

As all the values in the matrix are evaluated against the reliabilities (monotrait-monomethod diagonal), this value is an essential starting point from which to begin an MTMM analysis. There are many ways of calculating reliability and it is important to provide a procedure meant to serve non-experts with clear instructions and good tools. The most frequently cited reliability estimation is Cronbach's alpha (Cronbach, 1951). It is readily available within the menu options in SPSS and well known as a measure of internal consistency. Its greatest benefit is its familiarity and ease of calculation following a single administration of a test. Internal consistency refers to the degree to which items on a test are related to each other and ultimately to the same construct. However it has been criticized as a poor measure of reliability, representing at best a lower bound of reliability (Cortina, 1993; Cronbach & Shavelson, 2004; Sijtsma, 2009), and in some cases an overestimate of reliability (Raykov, 1998). Bliese (2000) has suggested another reliability estimation using intraclass correlations derived from within and between group mean squares. These can be calculated by conducting a one way ANOVA on the available data. The ICC (1) represents the reliability of a single assessment of a group-level property, or the degree to which one value in a group might fairly represent the group (Bliese, 2000). The ICC (2) provides an estimate of the group means (Bartko, 1976), (Bliese, 2000). There are several other reliability estimates that show promise as new standards for reliability estimates of educational data, particularly observational data (Sijtsma, 2009; Aiken, 1985; Hayes & K., 2007; Cronbach & Shavelson, 2004).

Reliabilities should be calculated for each new data set, but for the purposes of this paper, finding an appropriate calculation that is both simple and fairly represents the

data is an important part of the preliminary validation procedure being outlined. This paper will employ Cronbach's Alpha as the foundational reliability measure, to be used as the monotrait-monomethod values in the matrix. Given the default use of the Listwise function in SPSS to manage missing data, the alpha will allow all available data to be included in the analysis. The ICC will be calculated and discussed as a comparative reliability measure. The ICC measures require matched observations across cases (same number of observations per case) for comparison. The ICC will be used to evaluate inter-rater reliability because the multiple observations over two waves of data collection suggest the need for a reliability measure that evaluates the consistency of scores across observations. Cronbach alpha reliability measures will be used to assess the internal consistency of the scales within each of the data sources, following exploratory factor analyses. The following describes the intervention as it is ideally implemented. Past validation work is detailed as well.

The Youth Program Quality Intervention

After school programs, also referred to as Out-of-School-Time (OST) programs, represent a substantial portion of the large investment made in educational enterprises in the United States. In 2005, 40 percent of students in kindergarten through the eighth grade were participating in one or more non-parental, after school care arrangements (U.S. Department of Education, 2006). These programs are managed by a diverse group of nationally affiliated organizations, local governments and private institutions, and subsume a variety of content serving the full school-age population, including pre-school. They are both subsidized and fee-based. Such institutions provide critical support for working and low-income parents in terms of both enrichment and child care.

It is no surprise that defining and evaluating effectiveness or quality of such programs has been challenging (Scott-Little, Hamann, & Jurs, 2002) and their high rate of use suggests oversight, including assurances of quality, may be as critical for after school programs as for school-day programming.

The YPQI is a site-based program quality improvement model. It is an intervention based on an original program assessment metric first developed at the High Scope Foundation in 2005 by a group of youth program workers and teachers who were looking for a research-based professional development tool that put the evaluative and improvement power in the hands of the people doing the work, rather than outside evaluators. The measure was originally designed for after school programs, but it has since expanded its reach, both as a stand-alone measure and as a program intervention process, into regular school day programs. It is currently expanding its relevance for school day use via extended learning time initiatives, where the conventions of after school content are often blended with those of the standard school day. YPQI is currently being implemented in 3500 agency, school, and community-based settings in 95 networks, both within the United States and Internationally.

The YPQI is a structured sequence of performance measurements and performance feedbacks that program facilitators (typically, instructors or other direct service staff) can use to improve service delivery. These practice elements include the Youth Program Quality Assessment (Youth PQA). The YPQI is based in the Active-Participatory Method, a participatory learning approach developed at the High/Scope Foundation designed to support adult educators and caregivers of children, Pre K-early adulthood. According to the YPQI theory of change, incorporation of the quality

practices detailed in the Youth PQA and supported by the intervention will influence program norms to foster Instructional Quality. Instructional Quality is distinguished from Total Quality in the YPQI and the observational measure, Youth PQA, in terms of scores on the Youth PQA. Total Quality Score is an average score derived from all the items in each of the four domains (Safe Environment; Supportive Environment; Interaction; and Engagement). The Instructional Total Quality Score is the average of all the items in only three of the domains: Supportive Environment; Interaction; and Engagement. Because many of the items in the Safe Environment domain are regulated by entities outside the educational environment they are not included as part of the Instructional Total Quality Score (Smith & Hohmann, 2005; Smith, Pearson, Peck, Denault, & Sugar, 2009).

Once Instructional Quality is established as a program norm (as defined by high scores on the PQA tools), this higher degree of Instructional Quality will encourage youth cognitive and behavioral engagement (Akiva, T., Cortina, K., Eccles, J. & Smith, C., 2013; Naftzger, N, et al. 2014) and ultimately influence both academic and social outcomes for youth through a transfer of acquired skills.

The practice elements described by the YPQI are largely identified in the Youth PQA, a standardized observational measure. A School-Age PQA, intended for grades K-6 is also available to sites participating in YPQI. Many organizations use both in a given intervention process, depending on the ages served by the participating programs. Analyses in this paper will be confined to the data collected using the Youth PQA.

High fidelity to the YPQI model (See Table 2) includes at least one administration of a program self assessment using the Youth PQA (required); two observations for each participating site using the Youth PQA conducted by external assessors (external assessment is highly recommended, but not required for high fidelity to the YPQI. External assessment is conducted by certified external assessors trained using video-based recorded examples of instructional situations to 80% item-level perfect agreement with gold standard scores); a Planning with Data training, designed to train site-based teams to interpret the scores on the Youth PQA and to use the data to make Program Improvement Plans (PIP); instructional coaching, typically conducted by site managers with direct service staff in order to improve instructional practices, the YPQA is used as an instructional guideline; and some exposure to Youth Work methods, proprietary professional development courses designed to support improvement in the instructional practices identified by the YPQA. Although high fidelity is preferred, sites are not excluded from the intervention if they are not able to meet expectations for high fidelity.

The Youth PQA consists of four domains, and the version used in this secondary analysis consists of 18 scales and 60 items (The Youth PQA was updated in 2012 to include 63 items) (See Figure 1). Each item is scored on a three point Likert scale indicating the absence of a practice (1), the unintentional, occasional or partial availability of a practice (3) or the intentional and universal availability of that practice in terms of exposure to all session participants (5). The domains are organized in a pyramid form based on Abraham Maslow's hierarchy of needs (Maslow, 1954).

Maslow's hierarchy of needs is a theory of human motivation that argues that people first seek to meet basic needs for psychological and physical safety before they are able to address successively more complex needs. Once basic safety needs are addressed, he argued people seek to meet needs for involvement with community, also known as belonging, followed by recognition or esteem needs, and finally the highest level, self actualization, when individuals are able to realize their potential for emotional and intellectual growth. Maslow believed attainment of self actualization could be impeded by a failure to meet lower level needs.

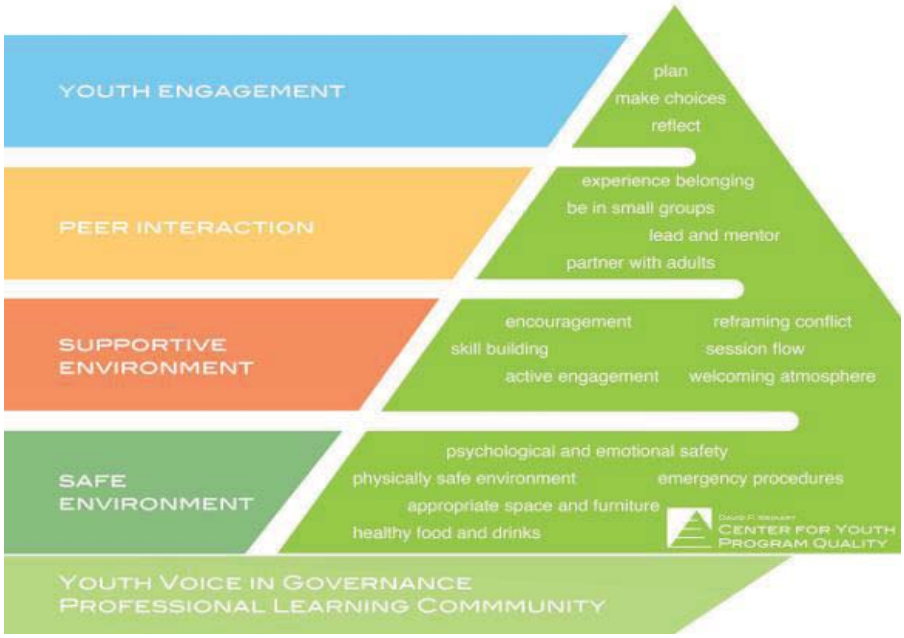

Figure 1 – Pyramid of Program Quality (Smith et al., 2012) Reprinted with permission from the author

Site officials are told by Weikart project managers (Weikart staff who coordinate services for the client/network) that high fidelity to the YPQI intervention model strengthens their ability to make inferences regarding the quality of the instructional

experience for youth in their programs. Among the elements of the YPQI, sites are expected to receive training on how to use the PQA measure. Youth PQA trained, site-based personnel are then expected to evaluate a sample of program offerings or sessions, using the Youth PQA. This evaluation is then used to start internal conversations about quality improvement, the expectation being that low scores on the measure indicate areas of needed improvement. These conversations are developed into a Program Improvement Plan (PIP) as site teams determine areas most in need of improvement, identified by low scoring Youth PQA items. PIPs are action plans detailing key areas of improvement fixed with specific steps and check in dates to monitor progress toward improvement goals. If sites are able to participate at an ideal level of fidelity, follow-up Youth PQA assessments using external assessors may be administered subsequent to the implementation of the PIP. In this way sites can determine if the areas that teams selected for improvement demonstrated measurable change on the Youth PQA scores. It is framed as a low-stakes initiation to the quality improvement process in part because site teams are so closely involved in the data gathering and interpretation process, but primarily because improvement in site-based practice provides the main incentive for low-scoring practices.

Table 2

YPQI Intervention Model: All Program Elements and Intervention Supports

Element	YPQI
Assessment	✓
PQA Kickoff: Introduction to YPQI and the Active-Participatory Method	Intervention support
Program self assessment at Baseline (Youth PQA): Site-based staff assess offerings (minimum of two per site) using the Youth PQA as a measure of quality	Required
External assessment at baseline: External raters trained to 80% inter-rater reliability are brought in to assess offerings. A minimum of two observations per site before the intervention (pre) and again following the intervention (post) is considered ideal.	Not required, but highly recommended
Planning	✓
Improvement Planning: Site teams use scores from the pre-intervention assessments to formulate action plans that target areas of needed improvement, identified by the Youth PQA.	Required
Instructional Coaching	✓
Site managers instruct staff on instructional improvement using the Youth PQA as a practice guide	Required
Youth Work Methods trainings	✓
Members of site teams attend trainings designed to improve instructional practices	Required
TA Coaching for site managers (focused on continuous improvement practices): Coaching from Weikart center available for site managers in how to support staff using the PQA as a model for improvement	Intervention Support
Quality Coaching Training for managers to coach staff (focused on improving instruction): training in the Observation-Reflection method of staff development	Intervention Support

Note: High fidelity to the model suggests adherence to the four YPQI elements: Assessment, Planning, Instructional Coaching, and Youth Work Methods (indicated by check marks)

The standard implementation of the YPQI typically includes staff and manager surveys as part of follow-up and forward planning for subsequent implementations of the intervention. While not a required part of YPQI implementation, they are mentioned here because survey data that include satisfaction information are included as part of the intervention data analyzed in this paper. The standard YPQI surveys are meant to assess both fidelity and satisfaction with the process from the point of view of direct service staff and their site-based administrators (managers). Sites in the YPQI validation study (from which data for this set of analyses were drawn) were also asked to complete follow-up surveys. Survey data was collected from participants at each level (administrative; direct service staff, youth participant) for the YPQI study. Point of service participants (direct service staff and youth) were surveyed immediately following the program offering (day of offering survey for both staff and youth) and again at the end of the programming cycle (program wide survey for both staff and youth). The program wide (culmination) survey was an extended survey meant to assess, in addition to the four main point of service constructs associated with the YPQI (Safety; Supportive Environment; Interaction; Engagement), culture and climate and general satisfaction. For staff, the satisfaction questions evaluated job satisfaction and existing quality improvement practices. For youth participants, the satisfaction questions referred to satisfaction with the instructor, the environment and the offering content. Youth survey questions also included motivation to attend and general emotional health questions.

Within the context of this analysis, the two data sources that will be used as the two different methods of measurement are the day of (observation) staff survey and the

YPQA observational measure. Both will be evaluated in terms of three traits associated with the previously defined trait constructs of a quality learning environment (Instructional Quality): Supportive Environment, Interaction, and Engagement. In this way, this paper attempts to address the presence or absence of the overarching construct of Quality, as it manifests in an instructional environment.

The Youth PQA was the subject of a validation study in 2005. Fifty-nine youth serving organizations serving the metropolitan areas surrounding Detroit and Grand Rapids were recruited for the study. Maximum available variation in terms of instructional content, youth population served, and location was sought for the sample. Once the organizations were selected, offerings within the organizations were selected based on the shared characteristics of regular meeting schedule over at least three months, with the same group of children for the same general purposes (Smith & Hohmann, 2005). In this way, the sample was selected to represent a wide range of programming environments and content, as is typical of after school programming. Following two waves of data collection, 356 Youth PQA ratings were completed and surveys were administered to 1,635 youth participants.

Among the findings of Smith and Hohmann (2005) was satisfactory evidence of inter-rater reliability. Inter-rater reliability was evaluated in two ways: intraclass correlations (ICC) were calculated within each rater pair and also across all pairs, and the percent perfect agreement between raters on each item. The ICCs were found to be within the acceptable level, demonstrating that there was greater variance across all pairs than within pairs. This suggested that pairs rating the same offering agreed about what they were seeing in terms of instructional practices during the offering. At the item

level, the average percent perfect agreement among rater pairs ranged from 48%-80% with the highest percentage agreement occurring in the Safe Environment domain (Smith & Hohmann, 2005). Inter-rater reliability was evaluated again in 2007. At that time, the researchers found that across 32 pairs of raters, there was 78% overall perfect agreement, at the item level, yielding an overall Kappa coefficient of .67 for the Youth PQA (Blazevski & Smith, 2007), indicating substantial overall agreement among raters (Landis & Koch, 1977).

Internal consistency was evaluated with Cronbach's alpha, yielding average alphas for the three subscales directly concerned with point of service or instructional quality: Supportive Environment (.85), Interaction (.68), and Engagement (.71), following the second wave of data collection (Smith & Hohmann, 2005). The three subscales cited are combined to form the Instructional Total Quality score. This score is made of three of the four observational domains in the Youth PQA. The Safe Environment domain is not included in this score, as the items are typically within the purview of state or federal guidelines (e.g. "there is a visible first aid kit") and as such demonstrate poor psychometric properties, specifically limited variance.

Principal component analyses conducted by Smith and Hohmann (2005) confirmed the structure of the subscales. These findings were replicated over two waves of data. Findings suggested that subscales were "related but empirically distinguishable constructs" (Smith & Hohmann, 2005, p.31). A correlation matrix conducted following the wave two data collection revealed that subscales associated with point of service quality or Instructional Total Quality Score were positively related to each other:

Table 3

YPQA: Subscale correlations

Subscale pairs	Correlation
Supportive Environment X Interaction	r=.61
Supportive Environment X Engagement	r=.61
Interaction X Engagement	r=.62

Note: *Correlational findings from Smith & Hohmann, 2005*

Smith and Hohmann (2005) believed the relatively strong relationship may have been due to the fact that the initial sample was not large enough. Another possibility was that the observational data subsumes information about youth and staff behaviors. In spite of the high correlation, repeated principal component analyses revealed distinguishable factors. The researchers pointed out that this instrument, which mixes items addressing both teacher behavior and child response, presents unique difficulties with respect to an easily interpreted set of measures. Inflated error variance in the item scores is noted as an understood cost of such measures, which are built on theory and consensus about best practices.

Concurrent validity was evaluated against the Youth Development Strategies, Inc. (YDSI) youth survey. Four subscales from the Youth Survey were selected as most closely aligned with the Youth PQA's four observational scales (Smith & Hohmann, 2005, p.17). Correlations indicated significant concurrent validity between aligned scales, following the second wave of data collection in the Interaction domain ($r=.44$, $p\leq.01$), the Engagement domain ($r=.32$, $p\leq.05$), and the Youth PQA Total Score (for scales I-IV; $r=.47$, $p\leq.01$). The Supportive Environment domain was not significant

(r=.29, p≤.1). Evidence of predictive validity was also found, but as this analysis is concerned with construct validity, the reader is referred to Smith and Hohmann (2005) for complete validity analysis.

Development of the Primary Data Set

The Youth Program Quality Intervention (YPQI) study was designed to study the impact of the intervention among diverse groups of after school settings. Ninety seven afterschool sites within five networks distributed over five states were initially chosen for the study. A network is a set of afterschool sites that share both geographic proximity (e.g., all within the same state) and policy context (e.g., 21st Century Community Learning Centers) (Smith et al., 2012, p.15). Networks were selected based on their ability to champion the work, deliver eligible sites, and support local delivery of essential YPQI elements, including YPQI training methods, and coaching and technical assistance. The study took place from 2006 to 2009. By the end of wave 3 (Spring 2008) 10 sites were dropped from the study due to program closure (9 sites) and refusal to participate (1 site). Site attrition analyses conducted as part of that study revealed that, when those 10 sites were dropped from the first wave of data collection (Spring 2007), intervention and control groups were not systematically different, although more of the dropped sites (7 of 10) were from the intervention group (Smith et al, 2012).

Diversity was an essential feature of this initial study, as diversity in program setting and instructional content is the most common feature of afterschool programs. It was anticipated that if impact might be demonstrated over a variety of settings, this would provide the most compelling evidence for generalizability (Smith et al., 2012). In

this spirit, sites were chosen to reflect the broad variety of afterschool settings including, Urban School District (Network A); State Department of Education (Network B); Independent Non-Profit (network C); School-Based Club (Network D); and State-Funded After School program (Network E). Programs within these networks also represented a wide range of programming goals for youth. Among the most common types of programming identified at sites during the first wave of data collection (Spring 2007) were Leadership (97% of sites provided some type of Leadership programming); Reading (96% of sites provided some type of Reading support programming); Physical Fitness (91% of sites); and Science (76% of sites) (see Appendix A, Table A1 for content offerings across baseline sites).

All networks except Network E successfully contributed all relevant data for all three years of the study. Difficulties with IRB approval prevented Network E from collecting youth survey data in a timely fashion during Wave 1 and some observational data from Wave 3 was lost by the data collection contractor due to a fire. The loss of Wave 3 observational data reduced impact analyses to four blocks and 68 sites, however effect sizes were large enough to overcome the threat of lost statistical power (Charles Smith, Personal Communication, February 1, 2014). As a result, Network E was excluded from impact analyses during the first study (Smith et al, 2012). Multiple sources of data were collected during Wave 1 (Spring 2007), Wave 2 (Fall 2007), Wave 3 (Spring 2008) and Wave 4 (Spring 2009). This secondary analysis uses data from Waves 1 & 3.

Chapter 3

Methodology

A secondary analysis of the data collected for the Youth Program Quality Intervention (YPQI) study (Smith et al, 2012) will be conducted, specifically to examine construct validity of instructional quality at the point of service. The presumed traits are Supportive Environment; Interaction; and Engagement, which were identified by previous validation work conducted by Smith and Hohmann (2005). The construct of Instructional quality will be examined with the multitrait-multimethod matrix (Campbell & Fiske, 1959) and evaluated using the I test (Sawilowsky, 2002).

Sampling

Two sources of data, and two methods of measurement will be used for this analysis. Each of these data sources were collected as part of the YPQI study (Smith et al, 2012), during the first and third data collection waves of the study (Spring 2007 and Spring 2008). The data sources will include: Waves 1 (N=255) and 3 (N=215) Day of (Observation) Staff Survey; and Waves 1 (N=190) and 3 (N=151) Youth PQA observations. Staff surveys were collected immediately following the program offering (program offerings are defined as a point of service setting where consistent groups of adults and youth meet over multiple sessions for the same learning purpose). Data for both the Treatment and the Control groups will be included to maximize the sample size. With two waves of data collection and two observations per offering, four moments in time will be represented for each offering.

The MTMM was intended to evaluate constructs via multiple traits and different methods. If sample sizes are adequate and the relevant scales contain items

representing the construct traits, the MTMM should be able to detect the presence of the construct across methods. Concerns might be raised with respect to the differing sample sizes that will be used in the present study. The observational measure will be evaluated with an initial sample of N=341 and correlated with the day of observation staff survey (N=470). Campbell and Fiske (1959) cite one study that evidenced “strong” “method variance” (Burwen & Campbell, 1957) which incorporated interview data (N=57) and a trait check list (N=155), indicating that they did not view unequal sample sizes as a problem for the MTMM (Campbell & Fiske, 1959, p. 88).The expectation for this study is that these sample sizes are large enough and contain adequate variance to mitigate small or uneven sample problems.

Constructs

The constructs to be evaluated using the MTMM will be determined for both data sources using principal components analysis with Varimax rotation. Factors will be evaluated and matched across methods. It is anticipated that each data source will render factors that closely approximate the three previously identified traits of Support, Interaction and Engagement (Smith & Hohmann, 2005). In the event this is not the case, new factors will be identified and compared in the MTMM across the methods. For a full listing of items selected for the present study, see Appendix A, Table A2.

One way in which this study will benefit from the previous validation work is in item selection. Only those items identified in the measure as the Instructional Total Quality Score (items in the Supportive; Interaction; and Engagement domains) will be incorporated in the analyses. Items in the Safe Environment domain will be omitted for

two reasons. First, these items tend to skew the total score higher than if the items were omitted. This is because items in the Safe Environment domain are often regulated by federal or state laws (e.g. is there a fire extinguisher visible) so not only do sites have little control of this domain (in terms of their ability to influence its quality score) but simple adherence to the law renders the highest score. The limited variance associated with this domain is the reason Instructional Total Quality Score is the primary quality measure used in the YPQI and also the reason only those items will be used for this analysis.

Following delineation of the trait constructs, which will be organized into comparable subscales within each data source, internal consistency reliabilities will be measured for each trait construct using the alpha statistic. Because the data used for the present study subsumes four different study conditions (Wave 1 and Wave 3 data and both treatment and control subjects), the reliability associated with group membership, ICC (1) (Bartko, 1976; Bleise, 2000), will also be measured for each trait construct.

Data Aggregation

The data will be obtained from archival SPSS files then cleaned and evaluated for missing values using an R matrix (Schafer & Graham, 2002; Rubin, 1976) in Excel. Overall missing data was not identified as a problem for any of the data sources in the original study.

Data for each of the three sources will be drawn from the first and the third waves of data collection and combined across the waves and both the treatment and control conditions in order to maximize the sample size.

Items on the staff survey administered during both Wave 1 and Wave 3 were scored on a three-point Likert scale reflecting the extent to which staff believed participating youth experienced the various learning conditions. For example, the statement “Youth were greeted within the first 15 minutes of the session” could be answered with the following choices: “1” = We did not do this today, “2” = This happened for some kids today, or “3” = This happened for almost all kids today. Seventeen items on the staff survey were scored on this scale. Three additional items asked the staff to estimate characteristics of the attending youth, including: percent “at risk” (“single parent household, low income, learning disability”), percent believed to be potentially successful in the program, and number of attending youth the staff believes to be afflicted with attention deficit disorder (ADD).

The staff surveys were completed by participants directly involved in the intervention. Items on the Youth PQA observational measure were scored by external observers (raters who did not participate in the intervention and were trained by the Weikart Center to 80% inter-rater reliability on the Youth PQA measure) who observed, on average, two separate offering sessions per site. A few sites collected only a single observation and a few sites collected three observations. Determinations about whether these sites will be retained for the present analyses will be made following missing data analyses on the data set for the present study. Average ratings of all observations collected for each site were created for summary analyses in the original study.

Observations were independent in so far as they were conducted for different staff and included different program content and different groups of participating youth. For both data collection methods (staff survey and observational measure) all items will be aggregated to the offering level across the four moments in time.

Design

Once the reliabilities (monotrait-monomethod) are established for each of the trait-method units, Pearson r correlations will be calculated for the monotrait-heteromethod values (validity diagonals), the heterotrait-monomethod triangles, the heterotrait-heteromethod triangles, per the instructions in Campbell and Fiske (1959). These matrices will then be evaluated for upward trend using the *I* statistic, per instructions in Sawilowsky (2002). Evidence of trend will suggest support for construct validity. These analyses will be looking for significance with a nominal alpha of .05. Evaluations of the *I* are made by counting inversions within the matrix. Original critical values in Sawilowsky (2002) identify 14 as the upper limit of inversions, given nominal alpha. The alpha reliability index will be employed in the MTMM and the ICCs will be calculated for comparison. The choice to compare different reliability indices was made to determine if different reliability estimations dramatically influence the correlation values within the matrix. It is reasonable to expect construct validation work to draw data from multiple populations. In such cases, the influence of group membership bears consideration along with the internal consistency of scales. The choices of reliability indices were made based on familiarity and current use. Other reliability indices may be more appropriate and these may be the subject of further study. The following hypotheses and guiding research questions will guide the study:

Hypotheses:

On the existence of the previously defined constructs:

a. Ho: Exploratory factor analyses using principal components analysis extraction will render constructs much as they have already been established by previous analyses.
b. Ha: PCA extraction will reveal different constructs than previously identified.

2. On demonstrating evidence of construct validity for the YPQI using the MTMM:

a. H_o: The 2X3 matrix of arrayed data will not present an upward trend using the I test with nominal alpha set at 0.05.
   H_a: The 2X3 matrix of arrayed data will present an upward trend using I test with nominal alpha set at 0.05.

Research questions for consideration:

a. Can a typical site-based administrator be expected to carry out the procedure outlined in this paper?
    a. Is the requisite equipment available to a typical site-based administrator?
    b. Does a general education graduate curriculum prepare a typical site-based administrator to conduct the analyses?
    c. Do the benefits of deepening one's engagement in the business of data collection and analysis justify the time required to conduct the analyses?

Analysis of the data will be conducted using available menu options in SPSS and Excel as much as is possible, in an effort to maintain necessary user friendliness. Results will be presented in tables, with narrative explanation of the steps, with the intent of providing replicable instructions for practical use.

Chapter 4

Results

Factor Analysis and Reliabilities

Initial cleaning of the data sets for both the YPQA and the Staff survey indicated a missing rate of less than 10%, except the K (Reframing Conflict) Scale on the YPQA. It had a missing rate of 17.5%, but it was retained for the analyses. Hence, the cleaned data sets included a total of 272 YPQA observations and 415 staff surveys.

The Safe Environment domain was omitted from the YPQA data due to item inconsistencies across waves. The three domains that constitute the Instructional Total Quality Score (ITS) were retained and represent the YPQA in these analyses. The YPQA data set was then analyzed for scale reliabilities, which were found to be consistent with earlier examination (Smith & Hohmann, 2005).

Reliabilities for the YPQA were recalculated using only the scales, as it was suggested by Smith (the principal investigator for the 2005 validation work; Smith and Hohmann, 2005) that analysis on the scales made more sense than on the items due to the formative nature of the traits. The reliability analysis, using the scales, determined an alpha of .69 for the Support domain. The Support domain included six scales (in this case, items) and 78 total observations. The missing data in the K scale (Reframing Conflict) caused 177 observations to be deleted by the Listwise function. When recalculated without the K scale, 254 of 272 YPQA observations were retained and alpha for the Support scale improved to .70, the gain slightly mitigated by the loss of the sixth scale. Due to the miniscule improvement following the elimination of the K scale, reliabilities used for the Support domain were calculated including the K scale. The

reliabilities calculated on the scales were considerably lower than those calculated on the items so the item reliabilities were selected for use in the matrices (see Table 4).

Table 4

YPQA – Comparison of Scale and Item Reliabilities

Domain	# Scales	Scale α	# Obs	# Items	Item α	# Obs
Support	6	.69	78	21	.78	52
Interaction	4	.60	248	12	.80	185
Engagement	3	.57	253	8	.75	267

Note: Use of Listwise function caused cases with missing data to be eliminated from the analyses

Factor analysis was conducted on the YPQA data using both PCA with Varimax rotation and Maximum Likelihood with Oblimin rotation. In both analyses, three factors were preselected for extraction. Both these analyses resulted in similar solutions, both supporting the existing scale construction (see Tables A4, A5).

PCA with Varimax rotation was also conducted on the Staff Day of survey. Three factors were forced (as in the PCA and ML for the YPQA data) in an attempt to match the previously identified three factors of the YPQA. Final scales for the staff survey are identified in Table 5. See also Table A2, for full item description. Reliabilities were calculated from the Waves 1 and 3, merged and cleaned, Staff Day of survey data file. Seventy-four total cases with 415 observations (surveys) were initially available for the staff survey using both the Wave 1 and Wave 3 data.

Table 5

Staff Day of Survey – Final scale details

Domain	Final Scale Items	# Items	# Obs	α
Support	d05, d06	2	402	.51
Interaction	d07,d08,d10,d11,d12	5	390	.57
Engagement	d09, d14,d15,d16,d17,d18,d19,d20	8	383	.74

Note: Use of Listwise function caused cases with missing data to be eliminated from the analyses

Intraclass Correlations (ICC)

All 74 cases (sites) were matched with two observations of the YPQA and two staff surveys for each wave. This ultimately resulted in the omission of 24 total sites (cases), leaving 63 total cases with matched data. The ICC analyses represent four time points, each time point including two PQA observations and two staff surveys. ICC values were calculated using a two-way mixed model (random people effects and fixed measures effects) based on absolute agreement between raters.

High values for the ICC (1) indicate, depending on the interpretation, a high proportion of variance attributable to group membership (Bryk & Raudenbush, 1984) or a highly reliable score on a given group attribute (high degree of agreement among individual scorers) (James, Demaree, & Wolf 1984). ICC (1) higher than .20 are rare (Bliese, 2000). Items in the sample which indicated high ICC (1) (≥ .20) are presented in Table 6.

Table 6

Item Descriptives – ICC(1)≥.20

Item	N	Range	Minimum	Maximum	Mean	SD	α	ICC(1)	ICC(2)
YPQA									
IIF1 All youth greeted in first 15 minutes	62	4.00	1.00	5.00	3.76	1.12	.54	.22	.54
IIJ3 Staff make frequent use of open-ended questions (staff ask open-ended questions throughout the activity)	63	4.00	1.00	5.00	2.64	1.11	.62	.29	.62
IIK3 In a conflict situation, adults ask youth what happened	46	4.00	1.00	5.00	2.84	1.46	.67	.33	.67
IIIL2 Youth exhibit predominantly inclusive relationships	63	3.00	2.00	5.00	3.79	.84	.57	.25	.57
IIIL3 Youth strongly identify with the program offering	63	3.00	2.00	5.00	3.82	.76	.53	.22	.53
Staff survey									
RISK What percentage of the kids in this session could be considered at risk	63	2.00	1.00	3.00	2.07	.72	.88	.66	.88

Calculated ICC (2) for items from both methods (YPQA and Staff Day of survey) were nearly identical to the calculated alphas for the items (see Table 6 and Table A2).

Following restructure for the ICC analyses, the internal consistency values showed the largest effect in the Support scale. In both the YPQA and the staff survey, all ICC (1) values were less than .1, some considerably less. In both the YPQA and the staff survey, domain reliabilities for the initial and restructured data sets were reasonably consistent with the ICC (2). See Tables 8 and 9.

MTMM

The first matrix was constructed using the initial merged data set (74 cases, 415 staff surveys, 272 YPQA observations). These included multiple observations in the same wave without regard to carefully matching or balancing observations across waves or measurement method. Reliabilities were first calculated as internal consistency (Cronbach Alpha) for each domain (Support, Interaction, and Engagement) within each method and within individual method data sets. Initial reliabilities for the YPQA rendered alphas comparable to those determined in the 2005 validation study, although the Interaction domain saw some improvement in the sample used for this analysis (see Table 7) (Smith & Hohmann, 2005).

Table 7

YPQA – Reliabilities: 2005 and Present Study

Domain	2005 W1 & W2 α (N=140)	Combined W 1 & W3 – Present study (N=272)
Support	.85	.78
Interaction	.68	.80
Engagement	.71	.75

As a matter of comparison, internal consistency reliabilities were recalculated for the domains, for both methods, using the restructured and combined data set. The four

moments in time associated with each item within each trait were combined across cases such that for each scale score in both methods there were four values. To maximize internal consistency, all item values were combined and included to calculate the scale reliabilities, such that the value for the Support trait in the staff survey, while the PCA determined it would only include two items, was calculated with eight total scores (one value on each item, from each of the four observations).

This method had variable effects on the trait reliabilities for the staff survey. Although internal consistency values decreased for both the Support and Interaction traits, the second alpha calculation (restructured data set) for Engagement showed slight improvement (see Tables 8 & 9).

Table 8

Changes in reliability values, following restructuring: Staff Day of survey

Staff Day of Survey	Initial Alpha – Initial combined data set	Second Alpha- Restructured, combined data set	ICC (1)	ICC (2)
Support	.51	.45	.09	.45
Interaction	.57	.54	.05	.53
Engagement	.74	.77	.08	.75

Note: ICCs were taken from the restructured data set

Table 9

Changes in reliability values, following restructuring: YPQA

YPQA	Initial Alpha – Initial combined data set	Second Alpha- Restructured, combined data set	ICC (1)	ICC (2)
Support	.78	.85	.07	.83
Interaction	.80	.82	.07	.79
Engagement	.75	.74	.07	.70

Note: ICCs were taken from the restructured data set

Once reliabilities for both the initial and restructured data sets had been calculated and compared, the initial reliabilities were selected for inclusion in the matrix. The initial combined YPQA and Staff Day of survey data were selected for calculation of the bivariate correlations that create the multitrait-multimethod matrix. All available data for both the staff survey and the YPQA were included in the first set of reliabilities and bivariate correlations. Correlations within the initial matrix, following Listwise deletion, used between n=264 and n=272 observations.

The MTMM – Four Matrixes

In the first matrix (see Table 10) two values in the validity diagonal indicated significant correlations, or correlations significantly different from zero, and also high enough to warrant further investigation, Campbell and Fiske's first criterion for convergent validity. The first was the correlation on the validity diagonal between the Interaction trait as measured by the YPQA and the Staff Day Of survey (monotrait-heteromethod). This correlation was significant at $p<.01$. The second significant correlation was the validity value for Engagement, which was significant at $p<.05$.

The Interaction correlation also meets the second Campbell and Fiske criterion, this one for discriminant validity, in that it is higher than the correlations in the heterotrait-heteromethod block, but it does not surpass that of the heterotrait-monomethod triangle (3^{rd} criterion for divergent validity). In other words the strength of the correlation between the scores of the Interaction scale as measured by the staff survey and the YPQA is not stronger than scores on different trait measures within a single testing method. Taken together, this suggests some evidence of both convergent

and discriminant validity indicated by the validity value associated with the Interaction trait.

The significant validity value associated with the Engagement trait, while significant at the $p<.05$ (meeting the first criterion), does not surpass all the other correlations in the associated heterotrait-heteromethod block (2nd criterion). It surpasses the correlations in the row, but not in the column, suggesting some level of discrimination. Nor does it distinguish itself in terms of being higher than the heterotrait-monomethod correllations, for either the YPQA or the staff survey (3rd criterion).

In terms of the fourth criterion, that "the same pattern of trait interrelationship be shown in all of the heterotrait triangles of both the monomethod and heteromethod blocks" – the Interaction trait distinguishes itself with the highest validity value. It is also the highest reliability for the YPQA method, but the mid-level reliability for the staff survey. In the staff survey monomethod block, Interaction shares the highest value correlation with Engagement. In the heteromethod block, Staff Interaction does not distinguish itself noticeably, sharing the lowest correlation in the lower heterotrait-heteromethod triangle with Engagement as measured by YPQA, and the mid-level correlation in the upper heterotrait-heteromethod triangle with Support as measured by YPQA. Interaction as measured by YPQA demonstrates both the highest reliability and shares the highest correlation in the YPQA heterotrait-monomethod triangle with YPQA Support. YPQA Interaction shares the mid-level correlation value in the lower heterotrait-heteromethod triangle with Staff Support. In the upper heterotrait-heteromethod triangle, the highest correlation is between YPQA Interaction and Staff Engagement. These results suggest some evidence of both convergent and

discriminant validity in terms of the Campbell and Fiske criteria. Additional analyses were then conducted with the *I* test (Sawilowsky, 2002) (see Table 11).

Table 10

Matrix 1 – df =264 - 406						
Staff				YPQA		
	Staff Supp	Staff Int	Staff Eng	YPQA Supp	YPQA Int	YPQA Eng
Staff S	(.51)					
Staff I	.185**	(.57)				
Staff E	.166*	.516**	(.74)			
YPQA S	*.020*	.088	.171*	(.78)		
YPQA I	.028	*.184***	.177*	.552**	(.80)	
YPQA Eng	.052	.-.004	*.131**	.476**	.494**	(.75)

*Note: * indicates significance at the .05 level. ** indicates significance at the .01 level*

I-Test

Table 11

I Test Values – Matrix 1:						
	Minimum		Median		Maximum	
Level	Value	I	Value	I	Value	I
Reliability	.51	0	.75	0	.80	0
Validity	.020	0	.131	0	.184	0
H-M	.166	2	.485	3	.552	4
H-H	-.004	0	.07	1	.177	3

Note. See Table 10. Total Inversions= 13. Pr [I ≤ 13] = 0.026 (see Table B2).

The I Test

In Matrix 1 (see Table 10), the *I* test revealed 13 inversions. This demonstrates an overall upward trend and is evidence of construct validity. This is consistent with the

findings in the initial correlation matrix based on the original Campbell and Fiske criteria (Table 10) and with previous construct validity findings for the YPQA (Smith & Hohmann, 2005). This finding also provides support for the construct validity of the YPQI in terms of the supportive relationship between methods within the intervention.

Further Investigation: Disattenuating the Correlations

In this case, dissattenuation (statistically removing the measurement error associated with the correlation) improved the originally significant validity values, such that the validity values for both Interaction and Engagement, assessed in the initial matrix as significant, then met the .001 significance threshold. Beyond this change in significance level little difference was found following dissattenuation. All values, except the heterotrait-heteromethod value for Staff Interaction and YPQA Engagement were improved slightly by disattenuation, however the overall pattern was unchanged from the initial matrix values. Significant validity values in the first matrix (Interaction and Engagement) remained significant in the disattenuated matrix, but failed to meet the third criterion, they did not surpass the heterotrait-monomethod values in the monomethod blocks (See Table 12). Trait relationships (4th criterion) remained the same.

Table 12

Matrix 1 – Disattenuated Correlations df=264-406						
Staff				YPQA		
	Supp	Int	Eng	Supp	Int	Eng
Staff S	(.51)					
Staff I	.343**	(.57)				
Staff E	.270**	.588**	(.74)			
YPQA S	.031	.132	.225**	(.78)		
YPQA I	.043	.272**	.230**	.699**	(.80)	
YPQA E	.084	-.0006	.175**	.623**	.638**	(.75)

*Note: * indicates significance at the .05 level. ** indicates significance at the .01 level.*

I Test: Disattenuated Matrix

The results of the *I* test for the disattenuated matrix were similarly unchanged. The total number of inversions (13) remained within the .05 significance cut-off. The dissattenuated correlations, however, did change the distribution of the inversions. The higher correlation values in the median and maximum spots on the H-M level caused an additional inversion for each value. The higher maximum value on the H-H level caused one less inversion for that value. (see Table 13).

Table 13

I Test Values – Matrix 1 – Disattenuated Correlations

	Minimum		Median		Maximum	
Level	Value	I	Value	I	Value	I
Reliability	.51	0	.75	0	.80	0
Validity	.031	0	.175	0	.272	0
H-M	.270	2	.605	4	.699	4
H-H	-.0006	0	.108	1	.230	2

Note. See Table 12. Total Inversions= 13. Pr [I ≤ 13] = 0.026 (see Table B2)

Restructured Data Matrix

The second matrix was created based on the recalculated reliabilites and bivariate correlations from the restructured data set (see Table 14). Overall, reliabilities were slightly lower. This was not surprising, given the smaller sample size. One significant correlation was found in the validity diagonal for the Interaction trait ($p<.01$). In terms of the heterotrait-monomethod correlations for the staff survey, only the correlation between the Interaction and Engagement traits was significant ($p<.01$). The heterotrait-monomethod correlations associated with the YPQA were all significant at the .01 level. Two correlations in the upper heterotrait-heteromethod triangle were also found to be significant. The pattern of correlations is very similar to the original matrix. All YPQA correlations are the highest; the highest correlation within the staff method is between the Interaction and Engagement traits, and within the H-H block, the validity value for Interaction is the highest, with the correlations between Staff Engagement & YPQA Support and Staff Engagement &YPQA Interaction, following in value and also significant. Again, the strongest traits appear to be Interaction, followed by Engagement, with the lowest correlations associated with the Support trait (domain).

Table 14

Matrix 2 – Restructured Data df= 230-251

	Staff			YPQA		
	Supp	Int	Eng	Supp	Int	Eng
Staff S	(.40)					
Staff I	.085	(.55)				
Staff E	.112	.463**	(.72)			
YPQA S	-.020	.121	.184**	(.78)		
YPQA I	.031	.216**	.166*	509**	(.77)	
YPQA E	.046	-.038	.042	.455**	.431**	(.71)

*Note: * indicates significance at the .05 level. ** indicates significance at the .01 level.*

The I-Test: Restructured Matrix

The values for the *I* changed very little (see Table 15). The final number of inversions was the same as the disattenuated matrix from the original data set, but here the smaller minimum reliability value caused an additional inversion for the median and maximum values on the heterotrait-heteromethod level.

Table 15

I Test Values – Matrix 2:

	Minimum		Median		Maximum	
Level	Value	I	Value	I	Value	I
Reliability	.40	0	.72	0	.78	0
Validity	-.020	0	.042	0	.216	0
H-M	.085	2	.443	3	.509	4
H-H	-.038	0	.044	2	.184	3

Note: See Table 14. Total Inversions = 14. Pr [I ≤ 14] = 0.037 (see Table B2)

Restructured Data – Disattenuated Correlations

When the correlations were disattenuated for measurement error, once again, as in both original (pre-disattenuation) matricies, the Interaction trait was the only significant validity value. All heterotrait-monomethod correlations, for both methods, were significant at the .01 level. One interesting change in the restructured matrix is the significance of the correlation between Staff Interaction & YPQA Support. A reappearance of the pattern from the earlier matricies would have shown a significant value in the Engagement validity. Here the removal of measurement error reveals a new significant correlation between Staff Interaction & YPQA Support, though it is also the lowest of the significant correlations in the H-H block (see Table 16).

Table 16

Matrix 2 – Restructured and Disattenuated df =230-251

Staff				YPQA		
	Staff Supp	Staff Int	Staff Eng	YPQA Supp	YPQA Int	YPQA Eng
Staff S	(.40)					
Staff I	.181**	(.55)				
Staff E	.209**	.736**	(.72)			
YPQA S	*-.035*	.185**	.245**	(.78)		
YPQA I	.055	*.332***	.223**	.657**	(.77)	
YPQA Eng	.086	-.06	*.058*	.611**	.583**	(.71)

*Note: * indicates significance at the .05 level. ** indicates significance at the .01 level*

I Test – Restructured and Disattenuated Data

Here the number of inversions in the *I* is affected by the lower minimum validity value. The validity value for Support has been consistently one of the lowest values in each of the matricies (surpassed only by the correlation between staff interaction and YPQA Engagement). This lower correlation adds to the inversions as well as the higher median value at the H-H level and in the case of this final *I* test, fails to meet the threshold for significance at the .05 level (see Table 17).

Table 17

I Test Values – Matrix 2: Restructured and Disattenuated

	Minimum		Median		Maximum	
Level	Value	I	Value	I	Value	I
Reliability	.40	0	.72	0	.78	0
Validity	-.035	0	.058	0	.332	0
H-M	.181	1	.597	4	.736	5
H-H	-.06	0	.135	2	.245	3

Note. See Table 16. Total Inversions= 15. Pr [$I \leq 15$] = 0.051 (see Table B2)

Chapter 5

Conclusions and Recommendations

Principal Components Analysis

In keeping with the method used in the original validation study (Smith & Hohmann, 2005) a popular and familiar exploratory approach that might likely be used by local or practice-level researchers, Principal Component factor Analysis (PCA) with Varimax rotation was conducted for both methods. The goals were to 1.) See if items on the YPQA generally replicated the same pattern identified in the first validation study (Smith & Hohmann, 2005) and 2.) See if the Staff Day of Survey would reflect similar patterns, specifically with respect to the domains of the Instructional Total Quality Score (Support; Interaction; and Engagement). These goals were related to the first hypothesis, that existing constructs were replicable, given the choice of PCA. Although the nature of PCA is more properly applied to large universes of items with the goal of finding the greater organizing concepts behind the data, PCA was chosen as a method that would likely be used given a practice level situation. Not only because it might be most accessible, but also because at that stage of analysis, organizing by principal components would likely be necessary. The expectation being that methods, measures or more specifically constructs drawn from practice level situations would probably be a collection of items or practices or other site-level indices that would require organization into principal components.

Another exploratory technique, Maximum Likelihood (ML), was also suggested. In the case of the data used for this investigation, it was reasoned that true factor analysis might be more appropriate. It was thought that ML might better serve the purpose of

identifying underlying constructs because the existing group of items was already limited, rather than attempting to reduce the data as in PCA. Additionally, because the data was known to be correlated, Oblimin rotation was also suggested (Charles Smith, Personal Communication, 6/18/2014). To address those concerns, comparative analysis was conducted on the YPQA with both PCA and ML (see Tables B2 and B3).

Little difference was found between the methods. Because a factor structure was already proposed, this investigation would have more properly employed Confirmatory Factor Analysis (CFA). However, CFA was not meant to be part of these analyses. This paper was intended to demonstrate construct validation through practically applicable means. The featured analyses are the MTMM and its evaluation by the *I* test. The procedures are intended to be simple, accessible and need not be overly precise. The process is meant to be applied by those at the local level with the intention that those closer to practice might be better equipped to contribute to the measurement conversation at the research level.

The Problem of the K Scale

The K scale or Reframing Conflict scale presents a persistent problem for evaluation of the YPQA. It is an important area of instruction to evaluate, especially in Out of School time programming, which tends to draw a disproportionate amount of youth in crisis. However, because it is dependent upon the existence of conflict in the instructional environment, items often go unobserved and create problems in terms of psychometric evaluation. It seems likely that the large amount of missing data in the K scale on the YPQA adversely affected the performance of the Support trait in the matrix. The following analyses were conducted as a means of comparison, to see what the analyses would have looked like given a simple means imputation and evaluation of the Support trait in the matrix with the omission of the K scale altogether. Upon reflection, means imputation really ended up hurting the results. In subsequent iterations of this method, it would be more appropriate to use either median of nearby points or linear interpolation – both easily accessible in the SPSS menu options, each offering a unique value for the missing value. In the case of the present data set, there were so few values it may not have made much difference, but that is still to be determined.

PCA for the YPQA was reevaluated given the K scale imputed with the overall mean of the observed K scale means. Only slight differences were identified in the loadings. Overall it was determined that inclusion of the K scale as it was presented no noteworthy differences with respect to the PCA analyses (see Tables 20 & 21).

Table 18

PCA with imputed K

Rotated Component Matrix

	Component		
	1	2	3
IIF Staff provide a welcoming atmosphere	.046	.067	.584
IIG Session flow is planned, presented, and paced for youth	.287	.192	.552
IIH Active Engagement	.577	.280	.241
III Staff support youth in building new skills	.303	.271	.572
IIJ Staff support youth with encouragement	.159	.367	.658
IIK Staff encourage youth to manage feelings and resolve conflicts appropriately	-.045	-.043	.610
IIIL Youth have opportunities to develop a sense of belonging	.315	.518	.377
IIIM Youth have opportunities to participate in small groups	-.042	.824	.054
IIIN Youth have opportunities to act as group facilitators	.239	.776	.100
IIIO Youth have opportunities to partner with adults	.557	-.030	.376
IVP Youth have opportunities to set goals and make plans	.626	.320	.116
IVQ Youth have opportunities to make choices based on their interests	.838	.041	-.099
IVR Youth have opportunities to reflect	.390	.488	.197

Table 19

PCA No K

Rotated Component Matrix

	Component		
	1	2	3
IIF Staff provide a welcoming atmosphere	.640	-.005	-.004
IIG Session flow is planned, presented, and paced for youth	.637	.233	.120
IIH Active Engagement	.354	.537	.222
III Staff support youth in building new skills	.716	.226	.163
IIJ Staff support youth with encouragement	.694	.125	.310
IIIL Youth have opportunities to develop a sense of belonging	.409	.304	.499
IIIM Youth have opportunities to participate in small groups	.056	-.021	.852
IIIN Youth have opportunities to act as group facilitators	.145	.244	.785
IIIO Youth have opportunities to partner with adults	.250	.602	.028
IVP Youth have opportunities to set goals and make plans	.166	.622	.310
IVQ Youth have opportunities to make choices based on their interests	-.063	.845	.053
IVR Youth have opportunities to reflect	.328	.347	.426

Imputation with a mean calculated from the average of the sample means for the K scale actually caused the K scale to negatively influence the performance of the Support domain. Here six values in the matrix decrease, and while the YPQA Support domain with the imputation appears to have caused the Support validity value to become significant, both the Interaction and Engagement validities which were significant when the K scale was included in spite of the missing data, have now lost significance. Five values, including the validity for the Support trait are improved with the imputation of the K values. The others are: YPQA Engagement & Staff Support; YPQA Interaction & Staff Support; YPQA Engagement & Staff Interaction; and YPQA Engagement & YPQA Support. Six values decreased following imputation: the reliability associated with the YPQA Support domain and the following correlations; YPQA Support & Staff Interaction; YPQA Support & Staff Engagement; Interaction Validity (down .084 and now not significant); YPQA Interaction & Staff Engagement (no longer significant); YPQA Interaction & YPQA Support; and the Engagement Validity (decreased by .114, no longer significant) (see Tables 20 & 21).

Table 20

Initial Matrix – df =264 - 406

Staff				YPQA		
	Staff Supp	Staff Int	Staff Eng	YPQA Supp	YPQA Int	YPQA Eng
Staff S	(.51)					
Staff I	.185**	(.57)				
Staff E	.166*	.516**	(.74)			
YPQA S	*.020*	.088	.171*	(.78)		
YPQA I	.028	*.184***	.177*	.552**	(.80)	
YPQA Eng	.052	.-.004	*.131**	.476**	.494**	(.75)

*Note: * indicates significance at the .05 level. ** indicates significance at the .01 level*

Table 21

Matrix: Imputed K (with mean) – df =263 - 407

Staff				YPQA		
	Staff Supp	Staff Int	Staff Eng	YPQA Supp	YPQA Int	YPQA Eng
Staff S	(.51)					
Staff I	.185**	(.57)				
Staff E	.166*	.516**	(.74)			
YPQA S	*.170***	.068	.068	(.69)		
YPQA I	.161**	*.100*	.100	.547**	(.80)	
YPQA Eng	.079	.001	*.016*	.480**	.494**	(.75)

*Note: * indicates significance at the .05 level. ** indicates significance at the .01 level*

When the K scale is removed from the YPQA Support domain, the difference is similar to imputation with the mean. Once again the reliability associated with the YPQA Support domain is decreased. The Support trait is improved with respect to its validity value, but the validities for both Interaction and Engagement disappear. The heterotrait-monomethod values for the YPQA are only modestly influenced with the omission of the K scale. Both the imputed matrix and the No K matrix indicate slight improvement for the Engagement & Support correlations, but the Interaction & Support correlations in both matricies are slightly decreased (see Tables 20 & 22).

Table 22

Matrix: No K – df =264 - 406						
Staff				YPQA		
	Staff Supp	Staff Int	Staff Eng	YPQA Supp	YPQA Int	YPQA Eng
Staff S	(.51)					
Staff I	.185**	(.57)				
Staff E	.166*	.516**	(.74)			
YPQA S	*.165***	.075	.078	(.71)		
YPQA I	.161**	*.100*	.100	.538**	(.80)	
YPQA Eng	.079	.001	*.016*	.496**	.494**	(.75)

*Note: * indicates significance at the .05 level. ** indicates significance at the .01 level*

Given the minor changes in the correlation values, most of which indicate compromised performance, it was expected the *I* test would be similarly affected. Results for the *I* test, for the imputed K matrix indicate slightly improved validity evidence, in so far as the number of inversions decreased by two, however significance is still within the .05 level (see Table 23).

Table 23

I Test Values – Imputed K Matrix

	Minimum		Median		Maximum	
Level	Value	I	Value	I	Value	I
Reliability	.51	0	.71	0	.80	0
Validity	.016	0	.100	0	.170	0
H-M	.166	2	.495	3	.547	4
H-H	.001	0	.078	1	.161	2

Note. See Table 21. Total Inversions= 11. Pr [I ≤ 11] = 0.012 (see Table B2)

When the K scale was omitted altogether, the *I* test gained one inversion, but significance remained the same ($p<.05$) (see Table 24).

Table 24

I Test Values – No K Matrix

	Minimum		Median		Maximum	
Level	Value	I	Value	I	Value	I
Reliability	.51	0	.73	0	.80	0
Validity	.016	0	.100	0	.165	0
H-M	.166	3	.495	3	.538	4
H-H	.001	0	.078	1	.166	1

Note. See Table 22. Total Inversions= 12. Pr [I ≤ 12] = 0.018 (see Table B2)

With respect to this data, inclusion of the K made little difference to the outcome. Because of this, the data was presented as it was found. Slight changes in the results of the *I* test suggest that given the severity of the missing data in a given sample, and there is every reason to believe missing data will always be a problem for education,

more sensitive imputation techniques than those presented here will be necessary. However, in the spirit of practical application, it is evident that even in cases of samples with high rates of missingness, preliminary validity evidence can be identified with the techniques outlined in this paper.

On the existence of the previously defined constructs

The hypothesis:

> Ho: Exploratory factor analyses using principal components analysis extraction will render constructs much as they have already been established by previous analyses.
> Ha: PCA extraction will reveal different constructs than previously identified.

It was expected that if the items and scales of the existing trait factors were strongly related, items and scale groups would reappear. This was in fact the case for the scale groups within the YPQA, and with respect to the first hypothesis, the findings suggest it is appropriate to accept the null. Items in the Staff Survey had not been previously subjected to factor analytic techniques and the results on the staff survey indicated that this analysis might have benefitted from a larger universe of initial items, for example the Support trait was represented by only two items following PCA and this had a negative influence on its reliability value which hurt its performance in the matrix. In spite of this, construct validity evidence was found in the heuristic analyses of the matrix and in the *I* but it seems likely that evidence might have been stronger given more sophisticated confirmatory techniques.

PCA was conducted on the scales of the YPQA, rather than the items, as the items have been found to be formative rather than reflective of scales. This is in contrast to the more common reflective perspective, where the scale or construct causes the indicators. Most measurement scales are evaluated as being reflective. Advocates of the formative perspective argue that this mischaracterization may result in model misspecification (Diamantopolis & Siguaw, 2006) and bias in terms of reliability estimations (Bollen, 1984). Because the original measure developers treat the YPQA as a formative measure (Smith et al, 2012 – STEM), that approach was also taken here, and scales rather than items on the YPQA, were submitted to PCA.

Multitrait Multimethod Matrix

Following the PCA, the data sets were combined and the first matrix was constructed. In terms of Campbell and Fiske's original evaluation criteria, the first consideration is the significance level of the values in the validity diagonal (see Table 25, values are indicated with italics). These values must be significantly different than zero and high enough to warrant further investigation. Table 18 presents the first matrix (before restructure and before disattenuation). The initial pattern of correlations was sustained through all but the final matrix which was constructed from the disattenuated, restructured data set. Only two of the correlations in the validity diagonal of the initial matrix meet Campbell and Fiske's first criterion (see Table 25).

Table 25

Matrix 1 – df =264 - 406						
Staff	YPQA					
	Staff Supp	Staff Int	Staff Eng	YPQA Supp	YPQA Int	YPQA Eng
Staff S	(.51)					
Staff I	**.185****	(.57)				
Staff E	**.166***	**.516****	(.74)			
YPQA S	*.020*	.088	.171*	(.78)		
YPQA I	.028	*.184***	.177*	**.552****	(.80)	
YPQA Eng	.052	.-.004	*.131**	**.476****	**.494****	(.75)

*Note: * indicates significance at the .05 level. ** indicates significance at the .01 level.*

The second and third criteria for continued analysis, is that values in the validity diagonal must be greater than other values in the same row and column. In this way the correlation between the same traits in different methods is not only higher than correlations between different traits in different methods (heterotrait-heteromethod (H-H) block), which is the second criteria; but it is also higher than correlations among different traits within the same method (heterotrait-monomethod (H-M) triangle), Campbell and Fiske's third criteria. For all of the matrices, the H-M correlations were highest, with the H-M correlations of the YPQA demonstrating the highest correlations of all. The final Campbell and Fiske criterion is that this pattern of relationships is consistent throughout the matrix. The most consistently strong correlations were associated with the Interaction domain, followed by the Engagement domain, with Support indicating the weakest relationships. This was true for the validity or M-H correlations. It was also true for both the YPQA and the staff survey, where Interaction

was indicated in the two highest values in the H-M correlations. With respect to the fourth criterion, there is some indication of repeated trait pattern in so far as Interaction appears to be the strongest influence in the relationships, among the domains. This finding echoes that of the 2005 work where Support also demonstrated the weakest correlations.

Table 26 presents all items of the Interaction domain of both methods. When taken in isolation, it seems clear that in both methods Interaction is, for the most part, intuitively obvious in terms of identification. For example, raters should be easily able to identify whether small groups happen in a given instructional setting, and it seems natural that if it is obvious to an external rater that groups exist, that it should also be clear to the staff that they created small groups, or more broadly, had opportunities for interaction with youth and youth had opportunities for interaction amongst themselves. No doubt easy recognition plays an important part in rater consistency. Also, as indicators in the Interaction domain of both measures feature multiple items related to grouping and person to person communication, it may be that these items are simply easier to connect to one another for both the measure/item developer and the analysis function, PCA in this case.

Table A2 presents the item level alpha-type reliabilities for each item. It is important to note all item reliabilities were taken from the restructured data set and each includes only four indicators so the reliabilities are generally low. Given that fact, items IIIL2 and IIIL3 were also among the highest reliabilities, .57 and .53, respectively. Interestingly, they were not the highest reliabilities. The highest alpha-type reliabilities were in the support domain, items IIJ3 (Instructor makes use of open-ended questions)

and IIK3 (In conflict situations, staff asks youth what happened), .62 and .67, respectively.

It may be that given the Support domain's greater number of items (N=21), that the strong reliability of some clearly identifiable items is diluted by the lesser clarity of others, for example item IIK1 (Every time there is a conflict involving strong feelings, staff ask about or acknowledge the feelings of the youth involved), α=.15.

Many of the clarity issues have been resolved as of the 2012 revision of both the YPQA and the SAPQA tools (School-Age version of the PQA tool). The highly problematic K scale has been dramatically simplified. Item IIK2, for example, has been reworked into two items; 1. Staff approach calmly; 2. Staff seeks youth input. It would be useful to revisit these analyses with updated data to see if these relationships persist.

Table 26

Interaction Items – YPQA and Staff Day of Survey
YPQA
IIIL1 Youth have structured opportunities to get to know each other
IIIL2 Youth exhibit predominantly inclusive relationships
IIIL3 Youth strongly identify with the program offering
IIIL4 Activities include structured opportunities to publically acknowledge the contributions of at least some youth
IIIM1 Session consists of activities carried out in at least 3 groupings
IIIM2 Staff use 2 or more ways to form small groups
IIIM3 Each small group has a purpose
IIIN1 All youth have multiple opportunities to practice group process skills
IIIN2 During activities all youth have opportunities to mentor an individual
IIIN3 During activities all youth have one or more opportunities to lead a group
IIIO1 Staff share control of most activities with youth
IIIO2 Staff always provide an explanation for expectations, guidelines, or directions given to youth
Staff Day of Survey
DSD07 Youth tried out new skills today or attempted higher levels of performance
DSD08 Youth worked as partners with staff on today's activity
DSD10 Youth had structured opportunities to get to know each other better
DSD11 Youth worked with youth collaboratively to complete todays activity
DSD12 Youth experienced 3 different groupings during the activity today

Note: Above includes all items in the Interaction domains of both methods. The two items highlighted in gray presented high (>.20) ICC(1) values, indicating a high degree of interrater reliability on those items for this sample.

Measurement Error

Overall, the results indicated some evidence of construct validity, but it was not particularly strong. In light of the limited analysis potential presented by the first matrix, it was decided that the analysis would benefit from the removal of measurement error where possible. Using Spearman's (1910) formula, this was done by disattenuating all the correlations of the matrix of the measurement error associated with the correlated scores. Spearman's formula is as follows:

$$R_{xy} = r_{xy} / \sqrt{(r_{xx} r_{yy})}$$

Limitations associated with the use of disattenuated correlations are that the disattenuated correlations are not directly comparable with uncorrected correlations and they are not suited to statistical hypothesis testing (Muchinsky, 1996). However, it can reveal correlations that may exist but are hidden due to measurement error. Because the process presented here is intended to be preliminary and accessible, the Spearman formula is especially useful, as it may help support arguments for validity by indicating potentially stronger correlations. This was the case for the present study.

Intraclass Correlations and Restructured Data Set

Intraclass correlations were taken using the menu options in SPSS. These provided a measure of reliability across raters (ICC (1)) and an estimate of the reliability of the group mean associated with a group characteristic (ICC (2)). The ICCs are important to the interpretation of the intervention's success, as typical YPQI implementation uses assessment by multiple raters to determine an overall quality rating. To calculate the ICCs across observations, the original data set had to be restructured such that each site was interpreted as a single case within which multiple observations on both the YPQA and the staff survey occurred. Ultimately the data sets were cleaned to create a restructured data set where each case had one observation per wave and per data source providing 4 total moments in time for each case to be compared on both data sources.

Some cases or sites had more observations on either the YPQA or the staff survey than other sites. One case had only two observations across both waves and methods (NY02 – two YPQA observations) and one case had ten staff surveys and four YPQA observations (MNc16). A minimum of one observation per method, per wave was necessary to conduct the analysis. Those cases which did not meet these criteria were omitted from this analysis. This resulted in five additional cases being omitted from the analysis, leaving n = 63. ICCs were calculated for each trait and in both methods. ICC (1) for all traits were less than .1, suggesting that multiple raters are necessary to provide an accurate assessment of site quality using the YPQA. This is consistent with earlier findings (Naftzger, Hallberg, & Yang, 2013) that multiple observations (ideally a minimum of three) are necessary to provide an accurate estimate of site quality based on the PQA measures. ICC (1) were also calculated for items across raters (See Table A2, Appendix A).

In cases of equal group sizes, the ICC (2) is a measure of the overall sample mean reliability. When ICC (2) were evaluated on the restructured data set, they were all within .01 of the alpha-type reliability estimates. Here, the group sizes were artificially cut to be matched, but given that they were even due to this matching, it is no surprise that alphas and ICC (2) were nearly identical in all cases. For future analyses it would wise to forego the Listwise deletion function and examine the ICC (2) with the original group sizes.

A MTMM was also constructed from the restructured data set. Overall, the correlations were lower, due to the smaller sample size. The interesting finding came from the disattenuated matrix where the Support domain actually got lower, following

disattenuation. Because it was assessed with a negative M-H value, following restructuring, that value was also lower following the disattenuation calculation (see Table 16). This was the case with all negative values following disattenuation, but only Support had a negative value in the M-H or validity diagonal.

The I Test

The *I* test is an easily calculated distribution-free test to support the evaluation of construct validity in the MTMM. The *I* test is an exact test of ascending trend of the a) heterotrait-heteromethod triangles, b) the heterotrait-monomethod triangles, c) the validity diagonal, and the d) reliability diagonals. Values for the *I* test are collected by taking the minimum, median, and the maximum values of each of the levels of the matrix. Starting from the bottom, the minimum, median, and maximum values of the heterotrait-heteromethod triangles are expected to have the lowest correlations. Each of these values is compared with those above them, more specifically with those values that are expected to be higher correlations. Each level of values is compared, one by one, with each of the values at each of the higher levels. Each incident of a lower value is considered an inversion and recorded.

For example, the maximum value (.177) on the heterotrait-heteromethod level in Table 11 has three inversions among the ascending comparisons: .166 on the heterotrait-monomethod level; .020 and .131 on the validity level. Each inversion interrupts the trend. No inversions would indicate the strongest evidence for construct validity. The cumulative distribution function (CDF) associated with the *I* indicates that 10 inversions is the upper limit for a significance of .01 and 14 inversions is the upper limit for a significance of .05 (see Table B2 for CDF). The *I* test indicates significance in

terms of the trend and whether its direction is overall (in spite of some inversions) ascending or the result of chance.

Limitations of the *I*

Perhaps the most evident limitation of the *I* is the data that is lost when the values are collapsed, as in the median values for each level, or when they are simply eliminated as in the unused values of the heterotrait-heteromethod triangles; the heterotrait-monomethod triangles; and the reliability diagonals (See Tables 10 and 11). Another related limitation is that while the *I* produces an evaluation of the matrix as a whole, it does not address the information in the specific trait-method units, nor does it permit analysis related to Campbell and Fiske's fourth criteria, the existence or nonexistence of repeated patterns among the levels of the matrix. In these ways it resembles other non-parametric analyses, specifically Hubert and Baker's non-parametric ANOVA, a variant of the generalized proximity function, in that it takes the average of the M-H correlations and it is evaluated by significance tests using a unique CDF (Hubert & Baker, 1978, 1979; c.f. Schmitt & Stults, 1986). In spite of these limitations the *I* is none the less an improvement over the heuristic criteria set forth by Campbell and Fiske and carries few of the opportunities for misspecification error of the more complicated procedures like Confirmatory Factor Analysis or the Direct Product Method.

Tables 11, 13, 15 and 17 describe the results of the *I* tests (Sawilowsky, 2002). The first matrix performed as well or better than all following, with the exception of the missing /imputed K matrices – neither of which changed the significance level. Notably, disattenuation added to the number of inversions for both the initial data set and the

restructured data set. In the first, disattenuation added only one inversion, in the restructured data set, two inversions were added following disattenuation. In both cases, negative H-H values (in the YPQA Engagement & Staff Interaction relationship) were lowered following disattenuation, but since all the other correlations were improved, this changed the relationships in the trend. For example, in Tables 13 and 17 changes in the median and maximum values at the H-M level increased the number of inversions following those values.

Construct validity for the YPQI using the MTMM

The hypothesis:

> H_o: The 2X3 matrix of arrayed data will not present an upward trend using the *I* test with nominal alpha set at 0.05.
> H_a: The 2X3 matrix of arrayed data will present an upward trend using *I* test with nominal alpha set at 0.05.

Evidence presented by the *I* test confirms upward trend, allowing the rejection of the null with respect to the second hypothesis. Heuristic analyses based on the original Campbell and Fiske criteria identified initial supporting evidence which was confirmed by the *I* test. While the evidence was not overwhelming, it was consistent and suggests further investigation is warranted. This confirms earlier validation work (Smith & Hohmann, 2005; Smith et al, 2012) and suggests that investigation with the current YPQI constructs, which include clarified items and scales, would provide stronger validity evidence.

Research Questions for Consideration

Can a typical site-based administrator be expected to carry out the procedure outlined in this paper?

Is the requisite equipment available to a typical site-based administrator?

Does a general education graduate curriculum prepare a typical site-based administrator to conduct the analyses?

Do the benefits of deepening one's engagement in the business of data collection and analysis justify the time required to conduct the analyses?

One of the central stated aims of this investigation was to create an approach to construct validation that might be used by site-based personnel with the goals of deepening local research knowledge and broadening practice-level potential for contributing to research. It is important to consider whether this is a reasonable expectation. Is it reasonable to think that administrators, practitioners, local evaluators might be interested in expanding their commitment to include measure development and/or validation?

It seems unlikely that in the face of shrinking budgets, which typically result in fewer staff to take on existing responsibilities, that site-based personnel will undertake additional tasks that require a lot of time. While the procedures outlined in this paper are simple in terms of methodology, the time it takes to translate instructional practice into a set of quantifiable actions and then become familiar enough with the basic concepts behind the procedures and the minimum requisite software to carry them out, is probably prohibitive. It is true that basic research methods which are part of most, but

not all, graduate instructional programs are probably enough of a foundation to understand these procedures, but many at the practice-level do not choose graduate school and those who do, but do not focus on methodology may have retained very little with respect to specialized software and research methodology.

However, in the hands of a motivated team that included a local evaluator to perform the analyses along with several practice-level contributors and perhaps an administrative level staff to support policy and other larger context considerations, these procedures might well provide a bridge between practice and research. These procedures might provide a path to quality improvement through the standardization of local practice. Educators who have lamented the fact that the ivory tower will never be able to understand how things are really done or who have developed methods that are validated only by agreement among colleagues might be empowered to contribute their local methods to the larger policy conversation.

The benefits of developing staff capacity around practice in terms of careful identification of what is truly successful, or not, about local practices cannot be underestimated. In terms of professional development, what could be more effective than intense self-reflection that leads to conversations about quality? This is, in essence what it means to dissect local instructional practice with the intention of identifying existing constructs. Even as a purely academic exercise, the process of parsing practice into recognizable constructs is beneficial in that it forces practitioners to make connections between actual practice and student skill development and in the case of YPQI, this also includes professional development. The procedures outlined in this paper make it possible to submit such musings to empirical evaluation. Educators can

contribute to the conversation about best practices both within and beyond the level of practice, and that is good for everyone.

APPENDIX A: CONTENT OFFERINGS ACROSS BASELINE SITES; ITEM DESCRIPTIVES; I TEST CDF; PCA

Table A1

Content Offerings across Baseline Sites

	Percentage of sites	Example program offereings
Leadership	97	Planning for team event, youth advisory board
Reading	96	Vowels, spelling
Life Skills	95	Race, culture
Art	93	Scrapbooking, clay
Physical Fitness	91	Walleyball, gym
Technology/Computers	90	Typing and navigating skills, video production
Math	89	Ratios, counts re: food drive donations
Community Service	89	Gifts to those in shelters
Sports	86	Basketball, baseball
Creative Writing	78	Journaling
Cooking	77	Recipies
Science	76	Laws of Motion
Dance	71	Hip Hop class
Music	71	History of pop music, guitar
Theater	69	Play rehearsal
Poetry	49	Poetry
Building/Shop	35	Robotics

SOURCE:*YPQI Study (Smith et al, 2012) reprinted with permission from the author*

Table A2

Item Descriptives

Item	N	Range	Minimum	Maximum	Mean	SD	α	ICC(1)	ICC(2)
IIF1 All youth greeted in first 15 minutes	62	4.00	1.00	5.00	3.76	1.12	.54	.22	.54
IIF2 During activities, staff mainly use warm tone of voice	63	1.00	4.00	5.00	4.72	.37	.16	.04	.15
IIF3 During activities, staff generally smile, use friendly gestures and make eye contact	63	2.00	3.00	5.00	4.62	.49	.45	.17	.44
IIG1 Staff start and end session within 10 minutes of scheduled time	63	2.00	3.00	5.00	4.53	.58	.32	.10	.32
IIG2 Staff have all materials and supplies ready to begin activities 9e.g. materials are gathered, set up, etc.)	63	2.67	2.33	5.00	4.56	.63	.48	.19	.48
IIG3 There are enough materials and supplies for all youth to begin activities	63	4.00	1.00	5.00	4.78	.59	.20	.06	.20
IIG4 Staff explain activities clearly (e.g. youth appear to understand directions; sequence of events and purpose are clear)	63	2.50	2.50	5.00	4.32	.62	.16	.04	.16
IIG5 There is an appropriate amount of time for all activities (e.g. youth do	63	3.00	2.00	5.00	4.14	.81	.16	.04	.16

not appear rushed, frustrated, bored or distracted; most youth finish activities)									
IIH1 The bulk of activities involve youth in engaging with (creating, combining, reforming) materials or ideas or improving a skill through guided practice	63	2.50	2.50	5.00	4.12	.74	.39	.14	.38
IIH2 The program activities lead (or will lead to in future sessions) to tangible products or performances that reflect ideas or designs of youth	63	3.50	1.50	5.00	3.60	.84	.04	.01	.04
IIH3 The activities provide all youth one of more opportunities to talk about (or otherwise communicate) what they are doing and what they are thinking about to others	63	3.50	1.50	5.00	3.86	.82	.31	.10	.31
IIH4 The activities balance concrete experiences involving materials, people, and projects (e.g. field trips, experiments,	63	2.00	2.50	4.50	3.61	.60	.12	.03	.12

interviews, service trips, creative writing) with abstract concepts (e.g. lectures, diagrams, formulas)									
III1 All youth are encouraged to try out new skills or attempt higher levels of performance	63	3.50	1.50	5.00	3.68	.92	.45	.17	.45
III2 All youth who try out new skills receive support from staff despite imperfect results	63	3.50	1.50	5.00	3.77	.85	.48	.18	.48
IIJ1 During activities staff are almost always actively involved with youth	63	2.50	2.50	5.00	4.52	.57	.33	.10	.32
IIJ2 Staff support at least some contributions or accomplishments of youth by specific non-evaluative language	63	3.00	2.00	5.00	3.21	.72	.39	.14	.39
IIJ3 Staff make frequent use of open-ended questions (staff ask open-ended questions throughout the activity)	63	4.00	1.00	5.00	2.64	1.11	.62	.29	.62
IIK1 Every time there is a conflict or an incident involving strong feelings, staff ask about or acknowlege the feelings of the youth involved	46	4.00	1.00	5.00	2.80	1.46	.15	.05	.17
IIK2 When strong feelings are	46	4.00	1.00	5.00	3.16	1.37	-.44	-.07	-.40

involved staff consistently help youth respond appropriately									
IIK3 In a conflict situation, adults ask youth what happened	46	4.00	1.00	5.00	2.84	1.46	.67	.33	.67
IIK4 As conflicts or incidents involving strong feelings occur, staff ask youth for possible solutions	45	4.00	1.00	5.00	2.43	1.41	.00	.00	.00
IIIL1 Youth have structured opportunities to get to know each other	63	3.50	1.50	5.00	3.14	.56	.32	.10	.32
IIIL2 Youth exhibit predominantly inclusive relationships	63	3.00	2.00	5.00	3.79	.84	.57	.25	.57
IIIL3 Youth strongly identify with the program offering	63	3.00	2.00	5.00	3.82	.76	.53	.22	.53
IIIL4 Activities include structured opportunities to publically acknowledge the contributions of at least some youth	63	3.00	1.00	4.00	2.68	.75	-.15	-.03	-.15
IIIM1 Session consists of activities carried out in at least 3 groupings	63	4.00	1.00	5.00	2.41	.93	.48	.19	.49
IIIM2 Staff use 2 or more ways to form small groups	63	3.50	1.00	4.50	2.01	.70	.29	.07	.25
IIIM3 Each small group has a purpose	63	3.50	1.00	4.50	2.70	.99	.09	.02	.09
IIIN1 All youth	63	4.00	1.00	5.00	3.33	.91	.13	.03	.12

have multiple opportunities to practice group process skills								
IIIN2 During activities all youth have opportunities to mentor an individual	63	3.00	1.00	4.00	1.91	.61	.14	.03	.13
IIIN3 During activities all youth have one or more opportunities to lead a group	63	2.50	1.00	3.50	1.64	.63	.35	.12	.36
IIIO1 Staff share control of most activities with youth	63	3.50	1.50	5.00	3.38	.86	.24	.07	.24
IIIO2 Staff always provide an explanation for expectations, guidelines, or directions given to youth	62	4.00	1.00	5.00	3.98	.84	.02	.00	.02
IVP1 Youth have multiple opportunities to make plans for projects	63	3.00	1.00	4.00	2.01	.86	.40	.15	.40
IVP2 In the course of planning projects, 2 or more planning strategies are used	63	3.50	1.00	4.50	1.69	.76	.46	.18	.47
IVQ1 All youth have opportunity to make at least one open-ended choice	63	4.00	1.00	5.00	3.07	.82	-.05	-.01	-.05
IVQ2 All youth have opportunity to make at least one open-ended process choice	63	4.00	1.00	5.00	3.35	1.01	.39	.14	.39
IVR1 All youth are engaged in	63	3.50	1.00	4.50	2.24	.89	.25	.07	.25

an intentional process of reflecting on what they are doing or have done									
IVR2 All youth are given the opportunity to reflect on their activities in one or more ways	63	3.00	1.00	4.00	1.77	.71	.43	.16	.42
IVR3 In the course of activity all youth have structured opportunities for presentation to the whole group	63	3.00	1.00	4.00	1.73	.81	.26	.08	.26
IVR4 Staff initiate structured opportunities for youth to give feedback	63	3.50	1.00	4.50	2.80	.79	.44	.16	.44
RISK What percentage of the kids in this session could be considered at risk	63	2.00	1.00	3.00	2.07	.72	.88	.66	.88
ADD_LD In your best estimate, how many kids in the session today have some type of attention deficit or learning disability	63	1.75	1.25	3.00	2.23	.40	.40	.15	.40
SUCCESS In your best estimate, what percentage of the kids are able to complete tasks consistently and learn successfully from the activities	63	1.67	1.33	3.00	2.15	.34	.11	.02	.10
DSD04 Youth were greeted by	63	.75	2.25	3.00	2.89	.17	.03	.00	.03

staff member in the forst 15 minutes									
DSD05 The kids had enough time to complete their activity or tasks for the day	63	1.00	2.00	3.00	2.81	.25	.30	.09	.29
DSD06 Youth understood the steps in completing todays activity	63	1.00	2.00	3.00	2.80	.24	.03	.00	.03
DSD07 Youth tried out new skills today or attempted higher levels of performance	63	2.00	1.00	3.00	2.25	.35	.05	.01	.05
DSD08 Youth worked as partners with staff on todays activity	63	1.50	1.50	3.00	2.35	.39	.07	.02	.07
DSD09 Youth were asked open-ended questions throughout the activity	63	1.50	1.50	3.00	2.23	.31	-.34	-.06	-.35
DSD10 Youth had structured opportunities to get to know each other better	63	1.75	1.25	3.00	2.22	.42	.34	.11	.34
DSD11 Youth worked with youth collaboratively to complete todays activity	63	1.25	1.75	3.00	2.53	.36	.43	.16	.43
DSD12 Youth experienced 3 different groupings during the activity today	63	2.00	1.00	3.00	1.95	.45	-.04	-.01	-.04
DSD13 Youth had the opportunity to take on leadership roles	63	2.00	1.00	3.00	1.89	.40	.24	.07	.24

today									
DSD14 Youth had multiple opportunities to practice group process skills	63	1.50	1.50	3.00	2.55	.34	.29	.09	.30
DSD15 Youth directed part of the session today	63	1.75	1.00	2.75	1.75	.38	.19	.05	.19
DSD16 Youth used 2 or more planning strategies for todays activity	63	2.00	1.00	3.00	1.82	.45	.40	.14	.40
DSD17 Youth had opportunities to make open-ended choices today	63	2.00	1.00	3.00	2.20	.42	.29	.09	.29
DSD18 Youth had the opportunity to mentor other youth today	63	1.75	1.00	2.75	1.93	.40	.24	.07	.23
DSD19 Youth reflected on what they did today	63	2.00	1.00	3.00	1.79	.47	.28	.08	.28
DSD20 Youth gave feedback about the activity today	63	1.75	1.25	3.00	2.05	.42	.17	.05	.17

Table A3

Cumulative Distribution Function (CDF) for the Number of Inversions (I) Test for Trend

I	CDF	I	CDF	I	CDF	I	CDF	I	CDF
0	0.00000271	11	0.01228896	22	0.26589286	33	0.81770563	44	0.99501894
1	0.00001082	12	0.01834416	23	0.31360119	34	0.85284904	45	0.99701299
2	0.00003517	13	0.02656926	24	0.36446699	35	0.88336580	46	0.99829004
3	0.00009470	14	0.03744318	25	0.41769751	36	0.90932900	47	0.99907197
4	0.00022186	15	0.05145292	26	0.47239719	37	0.93094426	48	0.99952922
5	0.00047078	16	0.06905574	27	0.52760281	38	0.94854708	49	0.99977814
6	0.00092803	17	0.09067100	28	0.58230248	39	0.96255682	50	0.99990530
7	0.00170996	18	0.1163420	29	0.63553300	40	0.97343074	51	0.99996483
8	0.00298701	19	0.14715097	30	0.68639880	41	0.98165584	52	0.99998918
9	0.00498106	20	0.18229437	31	0.73410714	42	0.98771104	53	0.99999729
10	0.00796807	21	0.22197240	32	0.77802759	43	0.99203193	54	1.00000000

Note: The CDF is produced by dividing the number of times each inversion occurs (0-54) by 369,600 (the total number of ways 12 values can be partitioned into 4 groups of 3) and summing the probabilities to the desired value. The CDF table was reprinted with permission by the author.

Table A4

Rotated Component Matrix – Principal Components Analysis with Varimax Rotation

	Component		
	1	2	3
IIF Staff provide a welcoming atmosphere			.511
IIG Session flow is planned, paced for youth	.433		
IIH Activities support active engagement	.596		
III Staff support youth in building new skills			.606
IIJ Staff support youth with encouragement			.767
IIK Staff encourage youth to resolve conflicts appropriately			.585
IIIL Youth have opps to develop a sense of belonging		.508	
IIIM Youth have opps to participate in small groups		.717	
IIIN Youth have opportunities to act as group facilitators		.571	
IIIO Youth have opportunities to partner with adults		.734	
IVP Youth have opps to set goals and make plans	.731		
IVQ Youth make choices based on their interests	.627		
IVR Youth have opportunities to reflect	.648		

Table A5

Pattern Matrix – Maximum Likelihood with Oblimin Rotation

	Factor		
	1	2	3
IIF Staff provide a welcoming atmosphere			.268
IIG Session flow is planned, presented, and paced for youth	.360		
IIH Activities support active engagement	.516		
III Staff support youth in building new skills	.503		.344
IIJ Staff support youth with encouragement	.308		.731
IIK Staff encourage youth to resolve conflicts appropriately	-.209	.412	.636
IIIL Youth have opps to develop a sense of belonging	.253	.267	.256
IIIM Youth have opps to participate in small groups		.312	
IIIN Youth have opportunities to act as group facilitators	.398	.361	
IIIO Youth have opportunities to partner with adults		.796	
IVP Youth have opportunities to set goals and make plans	.637		
IVQ Youth make choices based on their interests	.352	.427	-.373
IVR Youth have opps to reflect	.569		

APPENDIX B: FULL ITEM LISTING (YPQI MEASURES, INCLUSIVE): W1 & W3 YPQA; W1 & W3 YOUTH DAY OF SURVEY[1]

Table B1

Wave 1 Youth Program Quality Assessment (Youth PQA) Item Descriptives

	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
II.F.1 staff greet yth	158	4.04	1.44	-1.155	0.193	-0.113	0.384
II.F.2 staff warm/respectful	190	4.68	0.76	-2.18	0.176	3.696	0.351
II.F.3 staff smile/friendly	190	4.63	0.83	-2.081	0.176	3.516	0.351
II.G.1staff start/end 10 min sched	187	4.63	0.955	-2.589	0.178	6.009	0.354
II.G.2 material/supp ready	171	4.58	0.951	-2.224	0.186	4.301	0.369
II.G.3enough material/supp	158	4.87	0.538	-4.57	0.193	22.481	0.384
II.G.4staff explain actv clearly	188	4.34	1.22	-1.679	0.177	1.64	0.353
II.G.5 enough time for activts	190	4.19	1.348	-1.399	0.176	0.598	0.351
II.H.1activts transf/improve skills	189	4.19	1.269	-1.302	0.177	0.541	0.352
II.H.2 tangible prod reflect yth ideas	188	3.82	1.612	-0.879	0.177	-0.889	0.353
II.H.3 yth opps talk about doing/thnkng	189	3.8	1.44	-0.776	0.177	-0.695	0.352
II.H.4 balnce conc/abstract	188	3.78	1.139	-0.253	0.177	-0.776	0.353
II.I.1 encourage new skill	190	3.88	1.432	-0.887	0.176	-0.536	0.351
II.I.2 staff supprt new skills	180	3.91	1.359	-0.864	0.181	-0.429	0.36
II.J.1staff activ involved	190	4.51	1.083	-2.125	0.176	3.543	0.351
II.J.2 staff supp yth	190	3.43	1.319	-0.263	0.176	-0.741	0.351

[1] Vales in gray indicate noteworthy skewness

contribtns							
II.J.3 staff open-end questns	190	3.13	1.667	-0.119	0.176	-1.552	0.351
II.K.1 non-threat conf approach	68	3.15	1.704	-0.143	0.291	-1.617	0.574
II.K.2 conflict solut yth input	68	3.5	1.56	-0.474	0.291	-1.194	0.574
II.K.3 action/conseq relationship	66	3.12	1.75	-0.12	0.295	-1.7	0.582
II.K.4acknwldg/follow-up neg behav	65	2.94	1.657	0.058	0.297	-1.544	0.586
III.L.1 struct opps yth know each othr	188	3.32	1.217	-0.094	0.177	-0.397	0.353
III.L.2 yth exhibit incl relationships	189	3.98	1.265	-0.862	0.177	-0.288	0.352
III.L.3 yth ident w/program	190	3.97	1.177	-0.647	0.176	-0.533	0.351
III.L.4 structured opps acknwldgmt	190	3.03	1.677	-0.03	0.176	-1.58	0.351
III.M.1 incl mult group sizes	180	2.77	1.568	0.217	0.181	-1.343	0.36
III.M.2 2< ways to form sm grps	185	2.62	1.492	0.322	0.179	-1.141	0.355
III.M.3sm grp has purpose/all part	185	3.38	1.82	-0.384	0.179	-1.691	0.355
III.N.1 yth have opps pract grp sklls	190	3.38	1.716	-0.376	0.176	-1.543	0.351
III.N.2 yth have opps to mentor	189	2.54	1.51	0.405	0.177	-1.146	0.352
III.N.3 yth 1< opps lead grp	189	2.11	1.464	0.917	0.177	-0.557	0.352
III.O.1 staff/yth share contrl	190	3.65	1.579	-0.651	0.176	-1.096	0.351
III.O.2 staff expl expectations	153	4.23	1.15	-1.196	0.196	0.459	0.39
IV.P.1 yth have opps make plns	190	2.62	1.679	0.371	0.176	-1.485	0.351
IV.P.2 use 2< plan strats	190	2.32	1.599	0.697	0.176	-1.086	0.351
IV.Q.1 yth opps op-end content choice	190	3.2	1.574	-0.179	0.176	-1.362	0.351
IV.Q.2 yth opps op-end process choice	190	3.42	1.63	-0.406	0.176	-1.38	0.351
IV.R.1 intentional	189	2.58	1.698	0.42	0.177	-1.488	0.352

reflct proc							
IV.R.2 4yth opps to reflect	188	2.18	1.459	0.81	0.177	-0.69	0.353
IV.R.3 yth opps to present	189	2.23	1.743	0.839	0.177	-1.158	0.352
IV.R.4 struct. opps for feedback	186	2.84	1.28	0.072	0.178	-0.549	0.355
Valid N (listwise)	41						

Table B2

Wave 3 Youth Program Quality Assessment (Youth PQA) Item Descriptives

	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
IIF_i: All youth are greeted	124	4.00	1.498	-1.121	.217	-.287	.431
IIF_ii: staff use a	151	4.85	.522	-3.320	.197	9.145	.392

warm tone							
IIF_iii: Stf smile, eye contact.	151	4.76	.690	-2.856	.197	7.797	.392
IIF_iv: Emo. climate of the session is predominantly positive	151	4.50	1.012	-1.905	.197	2.854	.392
IIG_i: scheduled time.	149	4.54	.990	-2.115	.199	3.758	.395
IIG_ii: materials and supplies ready	133	4.65	.938	-2.779	.210	7.083	.417
IIG_iii: enough materials	131	4.79	.794	-3.899	.212	14.809	.420
IIG_iv: Staff explain all activities clearly	150	4.53	.967	-1.968	.198	3.158	.394
IIG_v:enough time	150	4.31	1.226	-1.582	.198	1.357	.394
IIH_i: Yth engaging with materials or ideas or improving a skill though guided practice.	151	4.36	1.163	-1.674	.197	1.771	.392
IIH_ii: tangible products or performances	151	3.82	1.637	-.890	.197	-.918	.392
IIH_iii: opportunities to talk about what they are doing or thinking	151	4.22	1.404	-1.509	.197	.734	.392
IIH_iv: The activities balance concrete/abstract	151	3.81	1.182	-.407	.197	-.683	.392
III_i: All youth are encouraged to try new skills	151	3.89	1.526	-.949	.197	-.627	.392
III_ii: staff allow youth to learn from	149	3.97	1.328	-.921	.199	-.287	.395

and correct their own mistakes							
III_iii: intentional opportunities for development of specific skills	150	3.76	1.744	-.825	.198	-1.178	.394
III_iv: Activities are appropriately challenging	150	4.17	1.208	-1.172	.198	.354	.394
IIJ_i: During activities, staff are almost always actively involved with youth	151	4.71	.745	-2.414	.197	5.075	.392
IIJ_ii: acknowledging what they've said or done with specific, nonevaluative lang	151	3.46	1.210	-.154	.197	-.493	.392
IIJ_iii: open-ended questions	151	2.81	1.703	.179	.197	-1.602	.392
IIK_i: acknowledge the feelings of the youth involved.	37	2.89	1.629	.102	.388	-1.480	.759
IIK_ii: staff encourage youth brainstorm possible solutions	36	3.17	1.540	-.146	.393	-1.261	.768
IIK_iii: In a conflict situation, adults ask the youth what happened.	37	3.00	1.826	.000	.388	-1.849	.759
IIK_iv: staff encourage yth to choose solution.	36	2.22	1.533	.813	.393	-.783	.768
IIIL_i: Youth have structured opportunities to get to know each other	146	3.42	1.056	.183	.201	-.057	.399

IIIL_ii: inclusive relationships	150	3.81	1.373	-.730	.198	-.614	.394
IIIL_iii: Youth strongly identify with the program offering	151	3.98	1.197	-.718	.197	-.439	.392
IIIL_iv: opportunities to publicly acknowledge	151	3.04	1.657	-.037	.197	-1.543	.392
IIIM_i: at least 3 groupings-full, small, or individual.	142	2.67	1.552	.314	.203	-1.274	.404
IIIM_ii: Staff use 2 or more ways to form small groups	146	2.21	1.379	.711	.201	-.646	.399
IIIM_iii: Each small group has a purpose	146	2.78	1.910	.222	.201	-1.883	.399
IIIN_i: multiple opportunities to practice group-process skills	150	3.79	1.604	-.833	.198	-.934	.394
IIIN_ii: one or more opportunities to mentor	151	1.82	1.271	1.290	.197	.518	.392
IIIN_iii: one or more opportunities to lead a group	151	1.74	1.257	1.481	.197	1.014	.392
IIIO_i: Staff share control of most activities	151	3.50	1.536	-.465	.197	-1.161	.392
IIIO_ii: explanation for expectations,	106	4.17	1.167	-1.072	.235	.178	.465
IIIO_iii: Staff talk with youth about their lives	150	3.01	1.711	-.013	.198	-1.638	.394
IVP_i: opportunities to make plans f	150	2.12	1.528	.939	.198	-.647	.394
IVP_ii: planning	150	1.93	1.384	1.168	.198	.032	.394

strategies							
IVP_iii: Staff support youths' planning by using and/or modeling effective planning strategies	150	2.17	1.608	.890	.198	-.866	.394
IVP_iv: All youth are encouraged to set project- or program-related goals	150	2.00	1.634	1.159	.198	-.486	.394
IVQ_i: open-ended content choice	149	3.34	1.777	-.337	.199	-1.658	.395
IVQ_ii: open-ended process choice	149	3.66	1.731	-.694	.199	-1.310	.395
IVQ_iii: Stf. encourage youth to take an activity in a new direction	147	3.08	1.637	-.076	.200	-1.503	.397
IVR_i: intentional process of reflecting	151	2.58	1.798	.432	.197	-1.632	.392
IVR_ii: MultiStratReflectionOpps	151	2.06	1.302	.842	.197	-.361	.392
IVR_iii: presentations	151	1.97	1.614	1.211	.197	-.354	.392
IVR_iv: Feedback	150	3.15	1.392	-.100	.198	-.915	.394

Table B3

Wave 1 Staff Survey: Day of Observation: Item Descriptives

	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error

RISK W1DSD01.Percentage of kids in session who are "at risk"	254	2.17	.852	-.332	.153	-1.546	.304
ADD_LD W1DSD02.Number of kids in session with ADD or other disability	254	2.37	.670	-.605	.153	-.681	.304
SUCCESS W1DSD03.Percentage of kids who are able to complete tasks and learn successfully from activities	251	2.08	.636	-.071	.154	-.532	.306
DSD04 W1DSD04.Kids were greeted within 15 minutes	255	2.89	.342	-3.026	.153	8.987	.304
DSD05 W1DSD05.Kids had time to complete the activity	255	2.84	.412	-2.475	.153	5.665	.304
DSD06 W1DSD06.Youth understood steps involved in completing activity	254	2.80	.423	-1.783	.153	2.044	.304
DSD07 W1DSD07.Youth tried out new skills or attempted hither levels of performance	252	2.32	.601	-.266	.153	-.629	.306
DSD08 W1DSD08.Youth worked as partners with staff	250	2.38	.679	-.630	.154	-.691	.307
DSD09 W1DSD09.Youth were asked open-	251	2.20	.675	-.264	.154	-.821	.306

ended questions throughout the activity							
DSD10 W1DSD10.Youth had opportunities to get to know each other	249	2.34	.750	-.644	.154	-.957	.307
DSD11 W1DSD11.Youth worked collaboratively with other youth	253	2.56	.612	-1.058	.153	.086	.305
DSD12 W1DSD12.Youth experienced three different groupings	251	2.06	.820	-.111	.154	-1.505	.306
DSD13 W1DSD13.Youth had the opportunity to take on leadership roles	250	1.90	.744	.156	.154	-1.174	.307
DSD14 W1DSD14.Youth had opportunities to practice group-process skills	253	2.53	.567	-.726	.153	-.489	.305
DSD15 W1DSD15.Youth directed part of the session	248	1.80	.762	.350	.155	-1.201	.308
DSD16 W1DSD16.Youth used two or more planning strategies	251	1.89	.795	.195	.154	-1.392	.306
DSD17 W1DSD17.Youth had opportunities to make open-ended choices	253	2.17	.781	-.307	.153	-1.298	.305
DSD18 W1DSD18.Youth had	254	1.98	.659	.017	.153	-.680	.304

the opportunity to mentor or teach other youth							
DSD19 W1DSD19.Youth reflected on what they did	248	1.83	.805	.309	.155	-1.393	.308
DSD20 W1DSD20.Youth gave feedback about the activity	250	2.10	.754	-.161	.154	-1.224	.307
Valid N (listwise)	232						

Table B4

Wave 3 Staff Survey: Day of Observation: Item Descriptives

	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
W3_RISK:In you best estimate, what percentage of the kids in the session could be considered "at risk" (single parent household, low income, learning disability, live in a high crime neighborhood,	212	2.14	.886	-.282	.167	-1.675	.333

etc.)?							
W3_ADD_LD: In your best estimate, how many kids in the session today may have some type of attention deficit or learning disability?	209	2.24	.700	-.369	.168	-.921	.335
W3_SUCCESS: In your best estimate, what percentage of the kids are able to complete tasks consistently and learn successfully from the activities that you provide?	212	2.23	.643	-.246	.167	-.670	.333
W3DSD04: Youth were greeted by a staff member within the first 15 minutes of the session	212	2.92	.289	-3.692	.167	13.936	.333
W3DSD05: The kids had enough time to complete their activity or tasks for the day.	210	2.82	.421	-2.233	.168	4.360	.334
W3DSD06: Youth understood the	214	2.80	.435	-2.007	.166	3.251	.331

steps involved in completing today's activity.							
W3DSD07: Youth tried out new skills today or attempted higher levels of performance.	212	2.21	.651	-.242	.167	-.703	.333
W3DSD08: Youth worked as partners with staff on today's activity.	211	2.45	.718	-.931	.167	-.484	.333
W3DSD09: Youth were asked open-ended questions throughout the activity (What do you think went wrong here?)	210	2.30	.686	-.478	.168	-.820	.334
W3DSD10: Youth had structured opportunities to get to know each other better.	209	2.25	.795	-.487	.168	-1.255	.335
W3DSD11: Youth worked collaboratively with other youth in order to complete today's activity.	213	2.53	.633	-1.015	.167	-.045	.332
W3DSD12: Youth	209	2.02	.799	-.043	.168	-1.431	.335

experienced 3 different groupings during the session today (full group, small group, pairs)							
W3DSD13: Youth had the opportunity to take on leadership roles today (leading a group session)	213	1.92	.702	.119	.167	-.957	.332
W3DSD14: Youth had multiple opportunities to practice group-process skills (actively listen, contribute ideas or actions to the group, do a task with others).	212	2.52	.619	-.937	.167	-.144	.333
W3DSD15: Youth directed part of the session today.	212	1.73	.734	.470	.167	-1.022	.333
W3DSD16: Youth used 2 or more planning strategies for today's project (brainstorming, idea webbing, backwards planning).	212	1.83	.797	.315	.167	-1.357	.333
W3DSD17:	211	2.23	.767	-.423	.167	-1.186	.333

Youth had opportunities to make open-ended choices today (topic selection, how to present results, order of activities)							
W3DSD18: Youth had the opportunity to mentor or teach other youth today.	212	1.86	.719	.210	.167	-1.042	.333
W3DSD19: Youth reflected on what they did today (writing in journals, sharing progress with the group)	214	1.82	.828	.352	.166	-1.454	.331
W3DSD20: Youth gave feedback (what they liked or disliked).	209	2.07	.784	-.127	.168	-1.361	.335
Valid N (listwise)	186						

Note: The first three questions, "Risk", "ADD", and "Success" were not included on the Wave 1 Staff survey

Table B5

Wave 1 Youth Survey: Day of Observation: Item Descriptive s

	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
W1YSD01_Staff seemed glad that I was here	792	.75	.435	-1.134	.087	-.717	.174
W1YSD02_I felt safe when I attended this session	788	.92	.267	-3.169	.087	8.061	.174
W1YSD03_I had an opportunity to take on responsibility	789	.68	.468	-.751	.087	-1.439	.174
W1YSD04_The staff encouraged me to try out new skills	791	.58	.493	-.337	.087	-1.892	.174
W1YSD05_I set goals for what I wanted to accomplish	788	.61	.489	-.432	.087	-1.818	.174
W1YSD06_I had enough time to complete what I was working on	790	.75	.431	-1.184	.087	-.599	.174
W1YSD07_I felt like the other kids in the session care about me	765	.62	.485	-.511	.088	-1.744	.177
W1YSD08_I had formal opportunities to get to know the other kids in the program	784	.64	.479	-.597	.087	-1.647	.174

W1YSD09_I was encouraged to reflect on the work I did	782	.66	.474	-.682	.087	-1.539	.175
W1YSD10_I had a chance to be "in charge" or lead others	779	.30	.460	.853	.088	-1.276	.175
W1YSD11_I felt like staff in today's session care about me	773	.79	.411	-1.392	.088	-.063	.176
W1YSD12_Other kids were willing to help me out	772	.66	.472	-.698	.088	-1.516	.176
W1YSD13_The staff challenged me to do my best	769	.63	.483	-.543	.088	-1.710	.176
W1YSD14_Staff members were prepared for today's session	770	.87	.333	-2.241	.088	3.030	.176
W1YSD15_I felt like I belonged to the group	761	.79	.406	-1.445	.089	.087	.177
W1YSD16_I had an opportunity to use my skills to help another kid	763	.45	.498	.198	.089	-1.966	.177
W1YSD17_I had an opportunity to present to the class	767	.28	.448	.995	.088	-1.013	.176
W1YSD18_I spent time planning how to complete a project	765	.40	.490	.420	.088	-1.828	.177
W1YSD19_I felt good about	770	3.94	1.228	-.994	.088	.023	.176

myself as a person							
W1YSD20_I enjoyed the session I attended	770	3.74	1.372	-.723	.088	-.732	.176
W1YSD21_The activities I did will help me get better at doing things I care about	767	3.51	1.446	-.494	.088	-1.073	.176
W1YSD22_The activities I did were challenging	769	2.36	1.423	.560	.088	-1.044	.176
W1YSD23_This session was interesting	767	3.37	1.404	-.347	.088	-1.123	.176
W1YSD24_I felt a sense of pride about what I was able to accomplish	767	3.37	1.408	-.339	.088	-1.154	.176
W1YSD25_I felt confident that I could do the activities in today's session	764	3.99	1.273	-1.063	.088	.022	.177
W1YSD26_How much choice did you have	760	3.01	1.418	-.039	.089	-1.191	.177
W1YSD27_How important was this activity for you	750	3.24	1.441	-.232	.089	-1.214	.178
W1YSD28_How much did you have to concentrate to complete the	755	3.17	1.426	-.189	.089	-1.198	.178

activity							
W1YSD29_Did you feel like you were using your skills	755	3.51	1.341	-.474	.089	-.869	.178
W1YSD30_Did you find yourself wishing you were doing something else	752	3.07	1.499	-.033	.089	-1.360	.178
W1YSD30_R (Reverse Scored)	752	2.93	1.499	.033	.089	-1.360	.178
Valid N (Listwise)	638						

Table B6

Wave 3 Youth Survey: Day of Observation: Item Descriptives

	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
W3_Q1: I feel like I belong at this program.	1155	3.21	.770	-.921	.072	.795	.144
W3_Q2: Kids worked together to solve problems.	1151	2.73	.842	-.356	.072	-.390	.144
W3_Q3: I feel like I matter at this program.	1142	3.07	.809	-.716	.072	.202	.145
W3YSD21: I got better at things I care	1144	3.05	.852	-.684	.072	-.091	.145

about.							
W3YSD23: I was interested in what we did.	1152	3.04	.920	-.783	.072	-.157	.144
W3YSD22: I was challenged in a good way.	1151	2.92	.896	-.567	.072	-.371	.144
W3YSD24: I felt a sense of pride about what I had accomplished.	1147	3.01	.864	-.670	.072	-.118	.144
W3YSD04: I tried to do things I have never done before.	1149	2.95	.915	-.573	.072	-.473	.144
W3YSD30: I wished I was doing something else.	1137	2.36	1.086	.238	.073	-1.227	.145
W3YSD30: I wished I was doing something else. (Reversed)	1137	2.64	1.086	-.238	.073	-1.227	.145
W3YSD26: I had a lot of choice about what we did.	1137	2.70	.924	-.344	.073	-.688	.145
W3YSD27: The activities were important to me.	1140	2.82	.909	-.426	.072	-.580	.145
W3YSD28: I	1148	2.73	.948	-.273	.072	-.839	.144

really had to concentrate to complete the activities.							
W3YSD29: I was using my skills.	1149	3.02	.900	-.750	.072	-.120	.144
W3_Q14: The activities were too easy.	1135	2.77	.967	-.183	.073	-1.030	.145
W3_Q14: The activities were too easy. Reversed	1135	2.23	.967	.183	.073	-1.030	.145
Valid N (listwise)	1032						

Table B7

Comparability of questions between wave 1 and wave 3 YPQI youth surveys

Wave 1 Youth Survey			Wave 3 Youth Survey		
W1YSD29	Did you feel like you were using your skills today?	"1" = Never "3" = Sometimes "5" = Almost all of the time	(Q13) W3YSD29	I was using my skills.	"1" = Yes "0" = No
W1YSD21	The activities I did today will help me get better at doing the things I care about.	"1" = Not at all true "3" = Somewhat true "5" = Very true	(Q4) W3YSD21	I got better at things I care about.	"1" = Yes "0" = No
W1YSD38	Over the past	"1" = Yes	W3YSD38	Over the past	"1" =

	month, have you been asked by staff to give your opinion on important program issues (selecting activities, deciding on a program schedule, arranging program space/furniture, hiring new staff)?	"0" = No		month, have you been asked by staff to give your opinion on important program issues (selecting activities, deciding on a program schedule, arranging program space/furniture, hiring new staff)?	Yes "0" = No
W1YSD39	Over the past month, has the staff encouraged you to become more involved in the youth program beyond just doing regular program activities (participate on an advisory panel, recruit other youth into the program)?	"1" = Yes "0" = No	W3YSD39	Over the past month, has the staff encouraged you to become more involved in the youth program beyond just doing regular program activities (participate on an advisory panel, recruit other youth into the program)?	"1" = Yes "0" = No
W1YSD15	When attending this program today, I felt like I belonged in the	"1" = Yes "0" = No	(Q1) W3_Q1	I feel like I belong at this program.	"1" = Yes "0" = No

	group				
W1YSD23	This session was interesting to me	“1” = Not at all true “3” = Somewhat true “5” = Very true	(Q5) W3YSD23	I was interested in what we did.	“1” = Yes “0” = No

REFERENCES

AERA, A., NCME. (1999). *The standards foreducational and psychological testing*: AERA.

Aiken, L.R. (1985). Three coeffiecients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement, 45*.

Akiva, T. (2012). The psychology of youth participation in organized activities.

Bagozzi, R., & Yi, Y. (1990). Assessing method variance in multitrait-multimethod matricies: The case of self reported affect and perceptions at work. *Working Paper: University of Michigan*.

Bagozzi, R., Yi, Y., & Phillips, L. (1991). Assessing construct validity in organizational research. *Administrative Science Quarterly, 36*, 421-458.

Bartko, J.J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin, 83*(5), 762-765.

Blazevski, J., & Smith, C. (2007). Inter-rater reliability on the youth program quality assessment *High/Scope Youth PQA Technical Report* (pp. 9).

Bliese, P. (Ed.). (2000). *Within-group agreement, non-independence, and reliability implications for data aggregation and analysis*: Jossey-Bass.

Bollen, K. A. (1984). Multiple indicators: internal consistency or no necessary relationship?. Quality and Quantity, 18(4), 377-385.

Browne, M.W. (1984). The decomposition of multitrait-multimethod matricies. *British Journal of Mathematical and Statistical Psychology, 37*(1), 1-21. doi: 10.1111/j.2044-8317.1984.tb00785.x

Bryk, A. S., & Raudenbush, S. W. Hierarchical linear models: Applications and data analysis methods, 1992.

Campbell, D., & Fiske, D. (1959). Convergent and discriminant validity by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2).

Campbell, D., & O'Connell, E. (1967). Method factors in multitrait-multimethod matrices: Multiplicative rather than additive. *Multivariate Behavioral Research, 2*(4), 409-426.

Campbell, D., & O'Connell, E. (1982). *Methods as diluting trait relationships rather than adding irrelevant systematic variance* (Vol. 12). San Francisco: Jossey-Bass.

Campbell, D., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*: Rand-McNally.

Cizek, G., Bowen, D., & Church, K. (2010). Sources of validity evidence for educational and psychological tests: Follow-up study. *Educational and Psychological Measurement, 70*(5), 732-743. doi: 10.1177/0013164410379323

Educational leadership policy standards 2008 (2008).

Cook, P. J., Dodge, K., Farkas, G., Fryer Jr, R. G., Guryan, J., Ludwig, J., ... & Steinberg, L. (2014). The (surprising) efficacy of academic and behavioral intervention with disadvantaged youth: Results from a randomized experiment in chicago (No. w19862). National Bureau of Economic Research.

Cortina, J. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98-104.

Cote, J. (1995). What causes estimation problems when analyzing mtmm data? , 2013, from http://www.acrwebsite.org/search/view-conference-proceedings.aspx?Id=9874

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*: Holt, Rinehart, and Winston.

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 22.

Cronbach, L., Rajaratnam, R., & Gleser, G. (1963). Theory of generalizability: A liberalization of reliability theory. *The British Journal of Statistical Psychology, XVI*(Part 2), 27.

Cronbach, L., & Shavelson, R. (2004). My current thoughts on coefficient alpha and successor procedures. *Educaltional and Psychological Measurement, 64*(3), 391-418.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Cross, A., Gottfredson, D., Wilson, D., Rorie, M., & Connell, N. (2010). Implementation quality and positive experiences in after-school programs. *American Journal of Community Psychology, 45*, 370-380.

Cuzzocrea, J. (2007). Robustness to non-independence and power of sawilowsky's i test for trend in construct validity. *Unpublished doctoral dissertation, Wayne State University, Detroit, MI*, 60.

Diamantopolous, A., Sarsttedt, M., Fuchs, C., Wilczynski, P., & Kaiser, S. (2012). Guidelines for choosing between multi-item and single-item scales for construct measurement: A predictive validity perspective. *Journal of the Academy of Marketing Science*. doi: 10.1007/s11747-011-0300-3

Diamantopoulos, A., & Siguaw, J. A. (2006). Formative versus reflective indicators in organizational measure development: a comparison and empirical illustration. British Journal of Management, 17(4), 263-282.

Durlak, J., & Weisberg, R. (2007). The impact of after-school programs that promote personal and social skills.

Erzberger, C., & Prein, G. (1997). Triangulation: Validity and empirically-based hypothesis construction. *Quality and Quantity, 31*, 141-154.

Fiske, D. (1982). Convergent-discriminant validation in measurements and research strategies. *New Directions for Methodology of Behavioral Science: Forms of Validity in Research, 12*, 77-92.

Fiske, D., & Campbell, D. (1992). Citations do not solve problems. *Psychological Bulletin, 112*(3), 393-395.

Fuchs, C., & Diamontopolous, A. (2009). Using single-item measures for construct measurement in management research conceptual issues and application guidelines. *Die Betriebswirtschaft*.

Fusarelli, L.D. (2004). The potential impact of the no child left behind act on equity and diversity in american education. *Educational Policy, 18*(1), 71-94. doi: 10.1177/0895904803260025

Garet, M., Porter, A., Desimone, L., Birman, B., & Kwang, S.Y. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Association, 38*(4), 915-945.

greatschools.org. (2013). Great schools.Org. greatschools.org http://www.greatschools.org/

Hall, D. (2013). A step forward or a step back: State accountability in the waiver era: The Education Trust.

Hayes, A., & K., K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77-89.

Heyneman, S.L., & Loxley, W. (1983). The effect of primary school quality on academic achievement across twenty-nine high and low income countries. *American Journal of Sociology, 88*(6), 1162-1194.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. Journal of applied psychology, 69(1), 85.

Kenny, D., & Kashy, D. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112*(1), 165-172.

Lance, C., Woehr, D., & Meade, A. (2007). Case study: A monte carlo investigation of assessment center construct validity. *Organizational Research Methods, 10*(3), 430-448. doi: 10.1177/1094428106289395

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics, 363-374.Lissitz, R., & Samuelson, K. (2007). A suggested change in terminology and emphasisregarding validity andeducation. *Educational Researcher, 36*(8), 437-438.

Loo, R. (2002). A caveat on using single-item versus multiple-item scales. *Journal of Managerial Psychology, 17*(1), 68-75.

Maas, C., Gerty, J.L.M., Lensvelt-Mulders, & Hox, J. (2009). A multitrait-multimethod analysis. *Methodology, 5*(3), 72-77. doi: DOI: 10.1027/1614-2241.5.3.72

Maslow, A. (1954). *Motivation and personality* (1st ed.): New York: Harper [1954].

Mathison, S. (1988). Why triangulate. *Educational Researcher, 17*. doi: 10.3102/0013189X017002013

Mayer, D. P. (2000). Monitoring school quality an indicators report. DIANE Publishing.

McCrae, R., Kurtz, J., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review, 15*(28), 28-50. doi: 10.1177/1088868310366253

Mehrens, W.A. (1997). The consequences of consequential validity. *Educational Measurement, 16*(2), 3.

Messick, S. (1989). Meaning and values in test validation: The science andethics of assessment. *Educational Researcher, 18*(2), 5-11.

Muchinsky, P. M. (1996). The correction for attenuation. Educational and Psychological Measurement, 56(1), 63-75.

Naftzger, N., Hallber,, K., & Yang, T. (2013). Exploring the relationship between qfterschool program quality and youth outcomes: A summary of findings from the prime ti,e of palm beach county quality improvement system. American Institutes for Reasearch. Naperville, IL.

Naftzger, N., Vinson, M., Liu, F., Zhu, B., & Foley, K. (2014). Washington 21st Century Community Learning Centers Program Evaluation: Year 2.

Nunnally, J. (1978). *Psychometric theory* (2nd ed.): McGraw-Hill.

OECD. (2013). Pisa 2012 results: What makes schools successful? Resources, policies and practices (Vol. IV, pp. 546): Organisation for Economic Cooperation and Development.

Pearson, K. (1895). *Mathematical contributions to the theory of evolution.--on a form of spurious correlation which may arise when indices are used in the measurement of organs.* Paper presented at the Proceedings of the Royal Society of London.

Phillipsen, L., Burchinal, M., Howes, C., & Cryer, D. (1997). The prediction of process quality from structural features of child care. *Early Childhood Research Quarterly, 12*, 281-303.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. Annual review of psychology, 63, 539-569.Popham, W.J. (1997). Consequential validity: Right concern- wrong concept. *Educational measurement: Issues and practice, 16*(2), 9-13.

Putka, D. J., Lance, C. E., Le, H., & McCloy, R. A. (2011). A Cautionary Note on Modeling Multitrait-Multirater Data Arising From Ill-Structured Measurement Designs. Organizational Research Methods, 14(3), 503-529.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. American Economic Review, 247-252.

Rubin, D. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592.

Sarstedt, M., & Wilczynski, P. (2009). More for less? A comparison ofsingle-item and multi-item measures. *Die Betriebswirtschaft, 69*(2).

Sawilowsky, S. (2002). A quick distribution-free test for trend that contributes evidence of construct validity. *Measurement and Evaluation in Counceling and Development, 35*, 11.

Sawilowsky, S. (2007). Construct validity. In N. Salkind (Ed.), *Encyclopedia of Measurement and Statistics* (Vol. 1, pp. 178-181). University of Kansas: Sage.

Schafer, J., & Graham, J. (2002). Missing data: Out view ofthe state ofthe art. *Psychological Methods, 7*(2), 147-177.

Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. Applied Psychological Measurement, 10(1), 1-22.

Schochet, P. (2009). Do typical rcts of education interventions have sufficient statistical power for linking impacts on teacher practice and student achievement outcomes? : Mathematica Policy Research.

Schools, A.A.P. (2012). Ann arbor public schools fy 2013/14 approved budget. http://www.a2schools.org/budget/files/2013-14approvedbudget.pdf

Schweinhart, L., & Weikart, D. (1997). The high/scope preschool curriculum comparison study through age 23. *Early Childhood Research Quarterly, 12*, 117-143.

Scott-Little, C., Hamann, M., & Jurs, S. (2002). Evaluations of afterschool programs: A meta-evaluation of methodologies and narrative synthesis of findings. *American Journal of Evaluation, 23*(4), 387-419.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*: Wadsworth.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of cronbach's alpha. *Psychometrika, 74*(1), 107-120. doi: 10.1007/S11336-008-9101-0

Sireci, S.G. (Ed.). (2009). *Packing and unpacking sources of validity evidence: History repeats itself again* (Vol. The concept of validity: Revisions, new directions and applications): Information Age

Smith, C., Akiva, T., Sugar, S., Devaney, T., Lo, Y., Frank, K., Peck, S., Cortina, K. (2012). Continuous quality improvement in afterschool settings: Impact findings from the youth program quality intervention study (pp. 169): The David P. Weikart Center for Youth Program Quality.

Smith, C., & Hohmann, C. (2005). Full findings from the youth pqa validation study (pp. 59): High/Scope.

Smith, C., Pearson, L., Peck, S., Denault, A.-S., & Sugar, S. (2009). *Managing for positive youth development: Linking management practices to instructional performances in out-of-school time organizations*. David P. Weikart Center for Youth Program Quality.

Smith, J. (2011). Reliability generalization: Lapsus linguae. 126.

Spearman, C. (1910). Correlation calculated from faulty data. British Journal of Psychology, 1904-1920, 3(3), 271-295.

Spielberger, J., Lockaby, T., Mayers, L., & Guterman, K. (2009). Ready for prime time: Implementing a formal afterschool quality improvement system by prime time palm beach county, inc. (pp. 135): Chapin Hall at the University of Chicago.

Statistics, N.C.f.E. (2012a). Projections of education statistics to 2021. Retrieved January 16, 2014, from Institute of Education Sciences http://nces.ed.gov/programs/projections/projections2021/tables/table_18.asp?referrer=list

Statistics, N.C.f.E. (2012b). Table 18. Actual and projected numbers for current expenditures and current expenditures per pupil in fall enrollment for public elementary and secondary education: School years 1996–97 through 2021–22. from Institute of Education Sciences http://nces.ed.gov/programs/projections/projections2021/tables/table_18.asp?referrer=list

Sternberg, R. (1992). Psychological bulletin's top ten hit parade. *Psychological Bulletin, 112*(3), 387-388.

Tay, L., Woo, S. E., & Vermunt, J. K. (2014). A Conceptual and Methodological Framework for Psychometric Isomorphism Validation of Multilevel Construct Measures. Organizational Research Methods, 17(1), 77-106.

U.S. Department of Education, N.C.f.E.S. (2006). The condition of education 2006.

UNESCO. (2011). Uis statistics in brief. Education all levels profile - united states of america 2011 USA. from UNESCO Institute for Statistics http://stats.uis.unesco.org/unesco/TableViewer/document.aspx?ReportId=121&IF_Language=eng&BR_Country=8400&BR_Region=40500

Education Week Professional Development Directory. Retrieved 1/16/2014, from http://pddirectory.edweek.org/

Woehr, D., Putka, D., & Bowler, M. (2011). An examination of g-theory methods for modeling multitrait-multimethod data: Clarifying links to construct validity and confirmatory factor analysis. *Organizational Research Methods, 15*(1), 134-161. doi: 10.1177/1094428111408616

ABSTRACT

BUILDING LOCAL SKILLS: THE MULTITRAIT-MULTIMETHOD MATRIX IN PRACTICE

by

ANNA C. GERSH

December 2014

Advisor: Dr. Shlomo Sawilowsky

Major: Educational Evaluation & Research

Degree: Doctor of Philosophy

The potential of expanding the evaluative skills of site-based practitioners is examined in a variety of educational enterprises by introducing a procedure to determine evidence for construct validity of measurement tools. The expectations of administrators of educational enterprises, including school day principals, administrators of after school and extended day programs, camps and other instructional settings to effectively collect and manage data is growing. Research skills are an important part of both accountability and improvement efforts which are frequently tied to funding. The multitrait-multimethod matrix (MTMM) (Campbell & Fiske, 1959) combined with the Sawilowsky I test (Sawilowsky, 2002) may provide a way for site-based administrators to develop a deeper understanding of the data for which they are responsible while increasing these administrators' usefulness as evaluators of measurement methods.

AUTOBIOGRAPHICAL STATEMENT

Anna C. Gersh MA, PhD (expected 12/14) www.linkedin.com/in/annagersh/
acgersh@gmail.com

Research and Evaluation Specialist with a Focus on Education

SKILLS

- SPSS, Qualtrics, Excel, Power Point, Word
- Technical Writing including technical reports, training guidebooks for professional development, feasibility studies, needs assessment
- Classroom Teaching including: developmental-advanced writing (college level); career readiness, writing, publishing, general math, photography (high-school); arts education and classroom support (K-12)

EXPERIENCE

7/11-Present **David P. Weikart Center for Youth Program Quality** Research Assistant

- Composed all written content and conducted data analysis for year-end evaluation reports for large-scale QIS
- Designed, evaluated and negotiated data analysis plans
- Authored guidebooks and online training content for instructor/coaching professional development
- Worked closely with project managers and network leads to ensure accurate and timely data collection and completion of deliverables (technical reports, executive summaries, folios)
- Articulated data analysis plans and procedures for non-technical audiences

4/10-12/12 **Wayne State University** Research Assistant

- Edited academic papers
- Conducted literature reviews
- Provided evaluation, data collection support, and content development for CORE grant program

9/02-9/10 **Washtenaw Community College** English Instructor

- Taught Developmental English; Argument; Technical Writing Classes

9/94-9/02 **Center for Occupational and Personalized Education (COPE)** Lead Teacher/Counselor

- Taught English, Journalism, Technical Writing, Career Readiness, Art
- Coordinated Summer School Program
- Created, implemented, and supervised school-wide portfolio evaluation program
- Instructed and supervised students in photography and darkroom technique
- Generated and maintained community partnerships
- Coordinated and supervised university volunteer program
- Wrote grant, awarded requested amount

EDUCATION

WAYNE STATE UNIVERSITY, Detroit, MI Doctoral Candidate/Educational Evaluation and Research Cognate: Educational Administration (expected December, 2014)

EASTERN MICHIGAN UNIVERSITY, Ypsilanti, MI MA Secondary Education/English Language & Literature (12/97); BS Majors: English Language & Literature/Telecommunications Minor: Psychology (12/88) Secondary Professional Teaching Certification in English & Psychology