Realist Evaluation of What Programs Work, for Whom and in What Circumstances: Utilizing Whole School Data to Improve Student Outcomes

Mansoor A. F. Kazi, and Yeongbin Kim, *University at Albany, State University of New York, US†

ABSTRACT

Research objective: Research indicates that the reform of the system of care and the use of initiatives such as wraparound is effective in improving mental health and functioning in school (Kutash, 2006; Reback, 2010; Goldenson, 2011). However, most studies have focused on at-risk groups rather than the total school populations. The purpose of this study was to help the school district undertake a 100% evaluation of school-based services, utilizing data on the entire school populations in a longitudinal study.

Methodology: This evaluation support was part of a $9million SAMHSA grant to develop a System of Care in New York State's Chautauqua County, and at no cost to the school district and the participating human service agencies. Based on the realist paradigm, data was regularly collected from a 100% school district population by Jamestown Public Schools (n = 5,535). Outcomes included academics, behavior, and attendance from 2011 and into the subsequent years. This article reports on the change in the academic achievement following programs of intervention in Jamestown Senior High School. Data analysis methods used with the school district included nonequivalent comparison group designs combined with logistic regression to investigate what interventions worked and for whom.

Results: It was found that, although Caucasian children and youth achieved better academic grades compared to Hispanic and African Americans, the binary logistic regression indicated that the actual predictors for improved academic achievement were the tutoring and mental health services as well as individualized education programs. It was found that those receiving these interventions had significantly improved their Grade

Point Average (GPA) scores when compared to those that had not received these interventions.

Conclusions: The interventions were having a positive impact on the children and youth in the school. This study indicates that a 100% evaluation with the entire population helps to better target and develop the interventions in a longitudinal, continuous evaluation.

INTRODUCTION

The main purpose of the evaluation was to provide real-time evidence based on the data already collected by the school district and the human service agencies providing the school-based programs consistent with the main purposes of the interventions, i.e., increasing student achievement and generating other positive health and wellbeing outcomes. The evaluators work with the school district and other partners to utilize the data that is already collected by the agencies and the school districts, to investigate what programs work and for whom, enabling the development of effective programs and/or the linkage of existing effective programs to the schools.

The evaluation research paradigm is based on realist evaluation as developed in Kazi (2003) and applied in Chautauqua County, NY, through a Substance Abuse and Mental Health Services Administration (SAMHSA) 's System of Care grant. In 2010, Chautauqua Tapestry (NY) received SAMHSA's Gold Award for Outstanding Local Evaluation, having used realist evaluation strategies in schools, mental health, youth justice and other human service agencies. These agencies routinely collect data that is typically not used for evaluation purposes. The 100% evaluation strategy utilizes a new approach to evidence-based practice based on the realist evaluation paradigm, with the central aim of investigating what interventions work and in what circumstances (Kazi, 2003). This approach essentially involves the systematic collection of data on 1) the student and school circumstances (e.g., demographic characteristics such as race and gender, special educational needs, lunch status, and parents' level of education); 2) the dosage, duration and frequency of the interventions received by the children and their families; and 3) the changes in the outcomes including school attendance, behavior and academic outcomes. This is a mixed methods approach, combining the traditions of epidemiology and effectiveness research in human services (Videka, 2003). This article focuses on the change in academic outcomes from the school year 2010-11 to the school year 2011-12.

BACKGROUND

This evaluation reports on findings from two school-based interventions In Jamestown Senior High School provided by two not-for-profit agencies. First, tutoring services (Safari academic services) consists of academic support partners provided to the school through Striders of Jamestown. The program is supported by a grant from the Sheldon Foundation and Jamestown Public Schools funding and is designed to provide academic tutoring support to students in need. The referrals are made by guidance counselors and teachers and the students are scheduled into the appropriate blocks of time within the school day. Tutoring is provided for all academic subjects and the tutors work with the classroom teachers to get the texts, assignments, projects and study guides that the students need to stay on top of their work. There is an after school component of the same service that is called Striders Tutors and it operates daily. Second, Family Service of Chautauqua Region, a Tapestry partner agency, provides a school-based therapy program. Through partnership with the agency and a United Way Grant, students are able to access mental health services in the school building. This assures that the students get to their appointments and it helps to return them to their classes quicker and with better strategies to help meet their needs in school.

School based interventions are often used to target specific student behavior, including behavior and socio-emotional problems and dealing with stress. Though the types and durations of interventions vary, the overall evaluation designs are similar, e.g., Comparison of the Rochester Resilience Project (Wyman et al., 2010); a mindfulness intervention (Mendelson, Greenberg, Dariotis, Gould, Rhoades, & Leaf, 2010), and the Pyramid Club Intervention (Ohl, Mitchell, Cassidy, & Fox, 2008). Goldenson (2011) examined the effectiveness of school-based mental health services focused on high-risk students. Specifically, this study examined "wrap around" services. According to Goldenson (2011), "wrap around" services emphasize the idea that the youth and family are central and that services should be individually tailored to the strengths or needs of the youth, rather than placing the youth in a program due to diagnostic criteria. Kutash et al. (2006) posited that the goal of this plan would be to develop coordinated, individualized treatment and/or service plans that would coordinate with or supplement Individualized Educational Programs (IEPs). In addition to personal benefits, results experienced by

individuals who participate in school-based services may also be experienced by those who interact with the individual, including peers, classmates, teachers, and family members (Goldenson, 2011).

These and other studies have indicated that school interventions can be utilized to make numerous improvements in the overall well being of students. Though they are implemented in different ways and last for a variety of durations, the evaluations of the interventions remain relatively similar. In order to evaluate interventions, most researchers utilize pre/posttest designs, looking at the initial baseline levels of the targeted behavior before and after the intervention is implemented. Additionally, post-intervention groups are often utilized to examine the effectiveness of the intervention.

However, most evaluations of interventions targeting school students do not use academic outcomes routinely and instead concentrate on the use of standardized measures or surveys. For example, Wyman et al. (2010) report on a school-based intervention for young people with behavioral and emotional problems. Apart from the use of data on suspensions, the program was evaluated using standardized behavioral measures and not the school outcomes. There are many examples of studies of school-based behavioral interventions that advocate increased use of evidence-based practices and evaluation but do not emphasize the routine use of successive marking period outcome data readily available in all schools (e.g., Gresham, 2004; Payne et al., 2006; Carswell et al., 2009; Filter & Horner, 2009). In their review of published research on school-based mental health programs, Rones & Hoagwood (2000) found that all of these evaluations typically ignored or underemphasized school-relevant outcomes.

Nevertheless, there is a growing trend to utilize the school data that already exists, e.g., Cook et al. (2007) found that friendship programs had a positive effect on school academic outcomes. In another study, Stormshak, Connell & Dishion (2009) used academic outcomes to evaluate family-centered intervention in schools with a sample of 998 sixth grade adolescents from three middle schools. They found that, compared to a control group not receiving the intervention, those who received the intervention improved their academic outcomes and attendance. Kazi (2014) evaluated school-based programs in

Jamestown, NY, using whole-school data, comparing academic outcomes of those receiving interventions with those who were not.

EVALUATION STRATEGY

The realist evaluation paradigm has been developed to investigate what programs work and in what circumstances, informed by the work of Pawson & Tilley (1997); Anastas (1999); Mark, Henry & Julnes (2000), and Kazi (2003, 2003a). The paradigm is defined as follows (Kazi, 2001: p.317):

'Realist evaluation seeks to evaluate practice within the realities of society. Practice takes place in an open system that consists of a constellation of interconnected structures, mechanisms and contexts. Realism aims to address all the significant variables involved in social work practice through a realist effectiveness cycle, which links the models of intervention with the circumstances in which practice takes place'.

At each cycle, a better approximation of reality is obtained with regard to what works, for whom and in what circumstances. Evidence from data gathering is used to adjust the practice to better meet the needs of the users and to generate the desired changes in the target problems that are being addressed by the practice intervention. This paradigm is inclusive, utilizing any research method or a combination of methods that can best reflect the realities of practice. There is growing recognition that the realist evaluation paradigm may address similar problems of evaluation also in related disciplines, e.g., in education (Batterbury & Hill, 2004; Priest & Walters, 2007) and in health (Clark et al., 2007). However, to date, there are only a few examples of realist evaluation, e.g., in social work (Kazi, 2003; Kazi et al. 2011, 2012), applied criminology (Tilley, 2004), and healthcare (Greenhalgh et al., 2009). Using a similar evaluation strategy, Bagdasaryan (2005) found that experimental and quasi-experimental studies produced mixed results with regard to the effectiveness of family preservation services in preventing out-of-home placements, and concluded that in order to better understand the relationships between family characteristics, service characteristics, and service utilization, it was necessary to inquire into the 'black box' of service delivery and examine the interactions between variables relevant for practice.

Many agencies use electronic databases to record practice data, and where the recording system is combined with the use of at least one reliable outcome measure, the opportunities are greater for integrating realist evaluation research into practice. In the evaluation of the school-based programs reported in this article, a data dump of the participating agencies' electronic databases, as well as data from the school district, was used within the agencies to investigate what works and in what circumstances.

The evaluation strategy focuses on both school attainment and, wherever possible, also on behavior in the classroom, and includes a longitudinal study for at least three years (and longer if permitted by reliable retrospective data, e.g., secondary data analysis of the two years prior to the grant). The findings from the earlier two years were reported in Kazi (2014); in this article, we are focusing on the academic years 2010-11 and 2011-12. The evaluation design is based on realist evaluation approach (Kazi, 2003) with the central aim of investigating with whom and in what circumstances the school-based interventions are more or less effective in improving academic outcomes for the school children.

All data from existing Management Information Systems (MIS) are analyzed at every marking period in partnership with the participating agencies and the school districts. These repeated findings inform practice to promote fidelity and to ensure that the services are targeted appropriately to meet the diverse needs of the children and families. The evaluators also help to develop these recording systems to identify other services that the families may be receiving, the extent of the inter-agency collaboration, and the impact of these services on the needs of the local community.

The research design is influenced by a combination of effectiveness research and epidemiology traditions. Where there is reliable retrospective data, it is possible to include the previous year as a retrospective baseline. Additionally, as the evaluators follow the children and young people from the baseline year and then into every successive marking period using the school outcomes, the school district will have other students with comparable demographics (e.g., lunch status or IEP status) who will not have received the school-based programs, and therefore will comprise natural nonequivalent comparison groups. In this way, evaluators can investigate the impact of the intervention in any two-year period, and also compare the outcomes between the school-based intervention school

cohort and the other non-intervention school children; and within the intervention school cohort by dividing children into groups according to the intensity of these interventions. Utilizing this secondary data, it is possible to examine patterns between the demographic variables, the dosage and frequency of the intervention and the school outcomes, to investigate the circumstances in which an intervention is more or less effective. The outcomes include school marking period assessments and state tests, attendance and where possible type of behavior incidents and the number of incidents, and other health data. Evaluators can also use the school academic grades for all children, and follow through each cohort of grade levels into the subsequent school years to investigate the impact of the intervention on not only behavior, but also academic performance and wellbeing.

The data can be entered in relation to each child. For example, when the outcomes are used repeatedly with each youth or teacher, it is a single-system design; when this data is put together for all classes receiving the same intervention, then this becomes a one-group pretest posttest design, and if there are children who have not received the intervention programs and/or have received other interventions, or if they receive different academic instructional services and/or other school based services and others have not, then it becomes a nonequivalent comparison group design.

As indicated by Videka (2003), the efficacy approach (testing hypotheses about the effects of an intervention) is in contrast to the epidemiological approach (explanation of the problem). However, blending these two research traditions would lead to an "explanation-oriented research approach that will result in the findings from clinical trials being more informative to the idiographic assessment-treatment approach, since it incorporates the study of individual and environmental characteristics with the study of service outcomes" (Videka, 2003, p. 178). As the research designs unfold naturally, data analysis methods can be applied to investigate the patterns between the client-specific factors, the intervention variables, and the outcomes. These methods will include the development of binary logistic regression (Kazi, 2003; Jaccard & Dodge, 2004) that identifies patterns in the data where multiple factors are influencing the outcome, and selects the main factor or factors responsible for the outcome, with a prediction of the odds of achieving a given outcome in particular circumstances. In this way, the outcome can be

linked to the potential causal factors without a control group. This evidence provides information about the effectiveness of the models of intervention in terms of what works, for whom and in what contexts. The evaluation helps the programs to better target their interventions, and to develop new strategies for the youth and teachers in the circumstances where the interventions are less successful. Comparison group designs can be used in which the participants are naturally assigned to program or comparison groups solely on the basis of a referral and acceptance, and the data analysis can potentially include causal analysis (Shadish, Cook & Campbell, 2001). All of these approaches to data analysis can be used at regular intervals whenever an outcome measure is repeated, e.g., every school academic term, and the findings can inform practice prospectively in each outcome measurement period.

In this process of evaluation, the evaluators are guided by Patton's utilization-focused evaluation (Patton, 2004), which includes changes in the culture, thinking and behavior of the organization receiving an evaluation that occur throughout the evaluation process. The critical issue here is that users can receive more from an evaluation than what is included in the analysis of findings. Of great importance in the utilization of the evaluation is that the data analysis is carried out in partnership with the stakeholders, and that they understand and have confidence in the information presented. Involving users in every step of the evaluation ensures that they take ownership of the findings and use them to develop and to better target the programs of intervention.

THE SAMPLE AND DEMOGRAPHIC

This evaluation utilized a 100% sample of all students in Jamestown Senior High School who had a grade point average in the academic year 2010-11 (n = 1299). The grade levels ranged from 9 to 12 (table 1), and the students were almost equally divided between the four grade levels. A total of 51.1% (n = 664) were male, 73.1% (n = 950, table 2) were white Caucasian, 59.9% (n = 778) were on free or reduced lunch, 12.3% (n = 160) had an Individualized Education Program, i.e. that the young person had a disability and needed special education; 17.6% (n = 228) had mothers who had achieved less than school graduation (i.e. did not go to college); and 12.4% (n = 161) had fathers who had not achieved school graduation (i.e., did not go to college).

Table 1. Frequencies of Grade Levels in Baseline (2010-11)

Grade Level	N	%
9th	349	26.9
10th	340	26.2
11th	309	23.8
12th	301	23.2
Total	1299	100.0

Table 2. Frequencies of Ethnicities in Baseline (2010-11)

Ethnicity	N	%
Asian	11	0.8
African-American	84	6.5
Hispanic	193	14.9
Native American	10	0.8
Mixed Race	51	3.9
White Caucasian	950	73.1
Total	1299	100.0

BASELINE OUTCOMES FOR SCHOOL YEAR 2010-11

Grade Point Average (GPA) was calculated by the School District for all Jamestown Senior High School students who were enrolled in the school year 2010-11. The mean GPA for the students (N = 1299) was 81.35, ranging from 9.00 to 99.45 (*SD* = 10.65) and 23.3% (n = 301) had achieved a GPA of less than 75.

Patterns in the data were explored between the baseline GPA achieved 75 or higher, and the demographic (or contextual) variables to investigate the significant differences in achievement. Using the Spearman test, it was found that:

- More females (80%, n = 508) had achieved a GPA of 75 or higher when compared with males (73.8%, n = 490), and there was a significant correlation between gender and whether a GPA of 75 or higher was achieved or not (r = .074, p <. 01, n = 1299, power = .99).

- Ethnicity was an important factor (Table 3). More White Caucasian students (80.7%, n = 998) had achieved 75 or higher when compared with all others (e.g., Hispanic 63.7%, n = 123) with the exception of Asian children who had all achieved 75 or higher, but they were a small minority (n = 11). When the database was divided between White Caucasian and all others, it was found that there was a significant correlation (80.7%, n = 767; r = .153, p < .01, n = 1299, power = .99).

Table 3. Disposition of Ethnicities Based on GPA for School Year 2010-11

Ethnicity Category		GPA		
		Less than 75	75 or Higher	Total
Asian	Count	0	11	11
	% Within Ethnicity	.0%	100.0%	00.0%
Black African-American	Count	32	52	84
	% Within Ethnicity	38.1%	61.9%	100.0%
Hispanic	Count	70	123	193
	% Within Ethnicity	36.3%	63.7%	100.0%
Native American	Count	6	4	10
	% Within Ethnicity	60.0%	40.0%	100.0%
Mixed Race	Count	10	41	51
	% Within Ethnicity	19.6%	80.4%	100.0%
White Caucasian	Count	183	767	950
	% Within Ethnicity	19.3%	80.7%	100.0%
Total	Count	301	998	1299
	% Within Ethnicity	23.2%	76.8%	100.0%

- Free or reduced lunch was also a significant factor. A higher proportion of those on paid lunches (90.8%, n = 473) had achieved 75 or higher than those on free or reduced lunch (67.5%, n = 525; r = .271, p < .01, n = 1299, power = .99).

- More students who did not have an IEP achieved 75 or higher (80%, n = 911) than those who had an IEP, and the two variables were strongly correlated (54.4%, n = 87; r = .199, p < .01, n = 1299, power = .99).

- Whether students' mothers had less than school graduation or a higher level of education was also a significant factor, with a smaller proportion with mothers with less than school graduation achieving 75 or higher (60.1%, n = 137) when compared to those with mothers who had school graduation or higher levels of education (82%, n = 801; r = .207, p < 01, n = 1205, power = .99).

- Fathers' level of education was also a significant factor. More students with fathers who had school graduation or more achieved 75 or higher (82.4%, n = 664) when compared with those with fathers who had school graduation or higher levels of education (62.7%, n = 101; r = .180, p < .01, n = 967, power = .99).

The above significant factors were entered in a forward-conditional binary logistic regression model with 'achieving GPA 75 or higher or not' as the dependent variable. It was found that those on free or reduced lunch were 3.2 times less likely to achieve 75 or higher than their fellow students on paid lunch; those with an IEP were 2.5 times less likely to achieve 75 or higher than the others with no IEP; students with mothers who had less than school graduation were 1.8 times less likely to achieve 75 or higher than those with mothers who had achieved school graduation or a higher level of education; and those who were not White Caucasian (i.e., Black African-American, Hispanic and others) were 1.9 times less likely to achieve 75 or higher than those who were White Caucasian.

Table 4. Binary Logistic Regression Results for Variables Predicting GPA 75 or Higher

Variables in Equation	B	SE	Wald	*df*	Sig	Exp(B)
Step 4 (c)						
Free or Reduced Lunch (1)	1.16	0.22	28.15	1	0.00	3.18
IEP (1)	0.90	0.24	14.68	1	0.00	2.47
Mothers' Education Level (1)	0.58	0.21	7.95	1	0.01	1.79
White	0.63	0.19	10.95	1	0.00	1.88
Constant	-2.55	0.18	205.17	1	0.00	0.78

Note. IEP = Individualized Educational Programs.

GPA ACHIEVED IN THE FOLLOWING SCHOOL YEAR 2011-12

When the students who were enrolled in 2010-11 were followed into the end of the next school year (N = 910, as grade level 12 had moved on), it was found (table 5) that mean GPA was reduced by 1.7% from the year 2010-11 to the year 2011-12, and that this small difference was significant {t(909) = 7.23, p < .01}.

Table 5. Mean GPA in Baseline (2010-11) and Following School Year (2011-12).

	Mean	N	SD	SE
GPA in 2010-11(Baseline)	81.84	910	10.24	0.34
GPA in 2011-12	80.14	910	10.56	0.35

However, when the individual students' change from the previous year into the next was taken into account, it was found that 35.7% (n = 325) had actually improved their GPA and the mean change for those who improved was 4.6 GPA points, as compared with a mean reduction of 5.2 GPA points for those who had not improved (64.3%, n = 585).

It was also found that a higher proportion (50%, n = 29) of the students receiving FSCR (N = 58) had improved when compared with those not receiving FSCR (34.7%, n = 296, improved; N = 852). Similarly, it was found that 62.7% (n = 32) of those receiving Safari (N = 51) had improved their GPA when compared with the students not receiving Safari (34.1%, n = 293, improved; N = 859). However, as the mean improvement for all students who had improved was 4.6 GPA points, the patterns in the data were investigated between those who had improved by 5 GPA points or more and those who had not, and the Spearman test indicated that the following were the significant factors:

- 22.1% (n = 25) of students who had an IEP (N = 113) improved their GPA by 5 points or more when compared with 10.3% (n = 82) of those that did not (N = 797) in the school year 2010-11; and the variable 'IEP or not' was significantly correlated to 'improved GPA by 5 points or more or not' (r = .121, p < .01, n = 910).

- Safari tutoring program or not. It was found that 31.4% (n = 16) of those who had received Safari (N = 51) in the school year 2010-2011 had improved their GPA by 5 points or more in the following year when compared with those who had not received Safari (10.6%, n = 91; N = 859), and that receiving Safari or not was correlated to GPA improved by 5 points (or more)or not (r = .148, p < .01, n = 910, power = .72).

- Family Service of Chautauqua Region (FSCR) mental health services received or not. It was found that 29.3% (n = 17) of those who received these services (N = 58) in the

school year 2010-11 improved their GPA by 5 points or more in the following year when compared with those who had not received FSCR services (10.6%, n = 90; r = .142, p = .019, n = 910, power = .72).

When the above significant factors were entered in a forward-conditional binary logistic regression model with the dependent variable GPA improved by at least 5 points or not (table 6), it was found that those who had received Safari were 2.8 times more likely to improve their GPA by 5 points or more when compared with those not receiving Safari; and the other predictors were receiving FSCR (3 times more likely) and having an IEP (2.3 times more likely).

Table 6. Binary Logistic Regression Results for Variables Predicting GPA improved by at least 5 points or not

Variables in Equation	B	SE	Wald	df	Sig	Exp(B)
Step 3 (c)						
Safari	1.03	0.34	9.18	1	0.00	2.79
FSCR	1.09	0.32	11.38	1	0.00	2.98
IEP	0.83	0.26	9.98	1	0.00	2.29
Constant	-.605	0.43	1.95	1	0.16	0.54

Note. FSCR= Family Service of Chautauqua Region. IEP = Individualized Educational Programs.

CONCLUSION

This evaluation demonstrates how the actual data collected by schools and human service agencies can be used for evaluation. In this example, the interventions were having a positive impact on the children and youth in Jamestown Senior High School. When a baseline year is selected with the academic achievement results for all students, the demographic factors such as lunch status, parents' level of education, ethnicity and special educational needs are correlated with achievement. However, when the students are followed through into the next academic year, the predictors for significant improvement become the individualized education programs, tutoring, and mental health services provided in the school. Findings at regular intervals from a 100% evaluation with the entire population help to better target and develop the interventions in a longitudinal, continuous evaluation. This approach to realist evaluation can be achieved by human services utilizing the data they routinely collect, and the findings can have a high utility for improving the

practice and impact of the human services. However, it is important to recognize that this is a partnership between academics and the human services, helping these services to utilize the data in a suitably de-identified form, and not the usual approach of the academics taking the data away and producing reports after practice has moved on. This article demonstrates that schools and human service agencies can use their data for evaluation and apply robust research methods to investigate their effectiveness at regular intervals, providing evidence in real-time to help to improve and to better target their programs of intervention.

REFERENCES

Anastas, J. W. (1999). Research Design for Social work and the Human Services. 2nd. Ed. New York: Columbia Press.

Bagdasaryan, S. (2005). Evaluating family preservation services: reframing the question of effectiveness. Children and Youth Services Review, 27, 615-635.

Batterbury, S. & Hill, S. (2004). Assessing the Impact of Higher Education on Regional Development: Using A Realist Approach for Policy Enhancement. Higher Education Management and Policy, 16(3), 36-52.

Carswell, S.B., Hanlon, T.E., O'Grady, K.E., Watts, A.M., & Pothong P. (2009). A preventive intervention program for urban African American youth attending an alternative education program: Background, implementation, and feasibility. *Education and Treatment of Children 32*(3), 445-469.

Clark, A. M., MacIntyre, P. D., & Cruickshank, J. (2007). A critical realist approach to understanding and evaluating heart health programmers. Health: An Interdisciplinary Journal for the Social Study of Health, Illness & Medicine, 11, 513-539.

Cook, T., Yingying, D., & Morgano, E. (2007). Friendship Influences During Early Adolescence: The Special Role of Friends' Grade Point Average. *Journal of Research on Adolescence (Blackwell Publishing Limited)*, *17*(2), 325-356.

Filter, K.J., & Horner, R.H. (2009). Function-based academic interventions for problem behavior. *Education and Treatment of Children, 32*(1), 1-19.

Goldenson, J. (2011). When there is no blueprint: The provision of mental health services in alternative school programs for suspended and expedited youth. *Child & Youth Services*, *32*(2): 108-123.

Goodman, R. (2001) 'Psychometric Properties of the Strengths and Difficulties Questionnaire (SDQ),' *Journal of the American Academy of Child and Adolescent Psychiatry*, 40, pp. 1337-1345.

Greenhalgh, T., Humphrey, C., Hughes, J., Macfarlane, F., Butler, C., & Pawson, R. (2009). How Do You Modernize a Health Service? A Realist Evaluation of Whole-Scale Transformation in London. Milbank Quarterly, 87, 391-416.

Gresham F.M. (2004). Current status and future directions of school-based behavioral interventions. *School Psychology Review 33*(3), 326-343. Jaccard,

Jaccard, J. & Dodge, T. (2004). 'Analyzing Contingent Effects in Regression Models,' in Hardy, M. & Bryman, A. (Eds.) Handbook of Data Analysis, Thousand Oaks, CA: Sage, p. 237-257.

Kazi, M. A. F. (2001) "Realist Evaluation," in Davies, M. (ed.) "The Blackwell Encyclopaedia of Social Work," Oxford: Blackwell.

Kazi, M.A. F. (2003). Realist Evaluation in Practice, London: Sage Publications.

Kazi, M.A.F. (2003a). Realist Evaluation for Practice.   British Journal of Social Work, 33(6), pp. 803-08.

Kazi, M. A. F. (2014) "Realist Evaluation and Effectiveness Research: An Example from School-based Interventions" in Kalliola, S. (ed.) 'Evaluation as a Tool for Research, Learning and Making Things Better,' Newcastle-Upon-Tyne, UK: Cambridge Scholars Publishing, pp. 31-44.

Kazi, M. A. F., Pagkos, B., & Milch, H. (2011). Realist Evaluation in Wraparound: A New Approach in Social Work Evidence-Based Practice. Research on Social Work Practice, 21(11), pp. 75-84.

Kazi, M. A. F., Frounfelker, S., Bartone, A., & Buchanan, P. G. (2012). Improving Outcomes for a Juvenile Justice Model Court: A Realist Evaluation. Juvenile and Family Court Journal, 63(3), pp. 37-54.

Kutash, K., Duchnowski, A. J., & Lynn, N. (2006). School-based mental health: An empirical guide for decision makers. Tampa, FL: University of South Florida, Research and Training Center for Children's Mental Health.

Mendelson, T., Greenberg, M. T., Dariotis, J.K., Gould, L.F., Rhoades, B.F & Leaf, PJ (2010). Feasibility and preliminary outcomes of a school-based mindfulness intervention for urban youth. . *Journal of Abnormal Child Psychology, 38:* 985-994.

Mark, M. M., Henry, G. T., & Julnes, G. (2000). Evaluation: An integrated framework for understanding, guiding and improving policies and programs. San Francisco, CA: Jossey-Bass.

Ohl, M., Mitchell, K., Cassidy, P., & Fox, T. (2008). The Pyramid Club primary school-based intervention: Evaluating the impact on children's social-emotional health. *Child and Adolescent Mental Health, 13* (3), 115-121. doi: 10.1111/j.1475-3588.2007.00476.x

Rones, M. & Hoagwood, K. (2000). School-based mental health services: a research review. *Clinical Child and Family Psychology Review*, 3: 223-241.

Wyman, P.A., Cross, W., Brown, C. H., Yu, Q., Tu, X., & Eberly, S. (2010). Intervention to strengthen emotional self-regulation in children with emerging mental health problems: Proximal impact on school behavior. *Journal of Abnormal Child Psychology, 38:* 707-720. doi: 10.1007/s10802-010-9398-x

Patton, M. Q. (2004). The roots of utilization-focused evaluation. In M. C. Alkin (Ed.) Evaluation roots: Tracing theorists' views and influences (pp. 276-292). Thousand Oaks, CA: Sage.

Pawson, R. & Tilley, N. (1997). Realistic evaluation. London: Sage. Programme: The theories of the programme and of its evaluations. Criminal Justice: International Journal of Policy and Practice, 4, 255-276.

Payne, A.A., Gottfredson, DC, & Gottfredson, GD (2006). School predictors of the intensity of implementation of school-based prevention programs: Results from a national study. *Prevention Science, 7,* 225-237.

Reback, R. (2010). Schools' mental health services and young children's emotions, behavior, and learning. *Journal of Policy Analysis and Management*, *29*(4), 698-725.

Stormshak, E.A., Connell, A., & Dishion, T.J. (2009). An adaptive approach to family-centered intervention in schools: Linking intervention engagement to academic outcomes in middle and high school. *Prevention Science, 10,* 221-235.

Tilley, N. (2004). Applying theory-driven evaluation to the British Crime Reduction. Vetere, A. (2001). Structural family therapy. Child Psychology and Psychiatry Review, 6(3), 133-139.

Videka, L. (2003). 'Accounting for variability in client, population, and setting characteristics: moderators of intervention effectiveness,' in Rosen, A. & Proctor, E. K. (eds.) Developing Practice Guidelines for Social Work Intervention, New York: Columbia University Press, p. 169-192.