DEVELOPMENTALLY APPROPRIATE EVALUATIONS

Developmentally Appropriate Evaluations: How Evaluation Practices Differ Across Age of Participants

Dana Linnell Wanzer

Claremont Graduate University

Abstract

It is well-known that meaningful differences in development exist among children, adolescents, and adults. However, to what extent do evaluators adapt their approach, designs, and methods to be responsive to a youth population? This study used an experimental simulation in which practicing evaluators designed an evaluation of a hypothetical tutoring program and were randomly assigned to the program serving children, adolescents, or young adults. Results indicate that evaluators were less likely to be participatory—both in the evaluation overall and in data collection—when the participants of the program were children than when they were adults. These findings suggest that evaluators, even those with experience conducting evaluations of youth programs, do not consider age as a meaningful cultural context and may need additional training in conducting developmentally appropriate evaluations.

Developmentally Appropriate Evaluations: How Evaluation Practices Differ Across Age of Participants

Children and adolescents differ from adults in their cognitive, socioemotional, and physical development. These differences require evaluators to adapt or modify their evaluation designs, methods, practices, and measures to be responsive to the age of participants (Berry, Menkes, & Bono, 2009; Christensen & Prout, 2002). In particular, methods of data collection, including surveys, interviews, focus groups, and observations, need to be adapted to accommodate the age and developmental stages of youth participants to ensure high validity and reliability of data produced. However, to what extent do evaluators adapt their designs according to the age of participants? Furthermore, do these adaptations relate to evaluators' experience with youth or training in youth development? This study will answer these questions through an experimental simulation study in which practicing evaluators will design and evaluation study of a hypothetical tutoring program with differing ages of program participants (i.e., children, adolescents, or adults).

Developmental Context

Many evaluators have described how the program context (e.g., age of program, type of program, feasibility) and the evaluation context (e.g., resources, stakeholder involvement, measurement tools) affect evaluation designs, methods, practices, and measures (e.g., Chen, 2004; Fitzpatrick, Sanders, & Worthen, 2004; Rossi, Freeman, & Lipsey, 1999; Weiss, 1998; Wholey, Hatry, & Newcomer, 1994). However, evaluators have only begun to examine how the developmental context—consisting of principles and theories of development, developmental domains, and age of participants—also affect evaluators design and conduct evaluations (Berry et al., 2009). In the American Evaluation Association (AEA)'s *Guiding Principles for Evaluators*, differences in age of participants require special consideration, stating “Evaluators

have the responsibility to understand and respect differences among participants, such as differences in their culture, religion, gender, disability, *age*, sexual orientation and ethnicity, and to account for potential implications of these differences when planning, conducting, analyzing, and reporting evaluations [*italics added*]" (AEA, 2004).

The difference between evaluation with adults and with children is more than simply a difference in *age*; rather, it is a difference in *culture* (Davis, 1998). Recognizing that children have a unique culture is a difficult endeavor: all adult evaluators have passed through childhood and subsequently believe they have a greater knowledge of children's culture than they actually do. As noted by Fine and Sandstrom, "Our [evaluators'] spatial proximity to children may lead us to believe that we are closer to them than we really are—only different in that (adults claim) children are still growing up ('developing') and are often wrong ('lack understanding')" (1988, p. 34). Thus, evaluation involving youth requires evaluators to demonstrate cultural competence in their evaluations.

The AEA also released a statement that described cultural competence in evaluation as "critical for the profession and for the greater good of society" (AEA, 2011). Culturally competent practices in evaluation include acknowledging the complexity of cultural identity, recognizing the dynamics of status and power, recognizing and eliminating bias in language, and employing culturally appropriate methods (AEA, 2011). In particular, the use of culturally appropriate methods requires careful consideration of the methods and tools used for collection, analysis, interpretation, and dissemination of data to ensure reliable data and valid inferences (AEA, 2011; Choinard & Cousins, 2009). This statement indicates the need for evaluators to employ developmentally appropriate practice to demonstrate cultural competence in the context of youth programs.

Just as cultural competence requires inclusion of cultural expertise on the evaluation team

(Chouinard & Cousins, 2009), developmentally appropriate practice requires inclusion of developmental expertise on the evaluation team (Bialosiewicz, 2013). Developmental expertise can help evaluators serving youth programs to create evaluations aligned with the developmental context (Eide & Winger, 2005; Rapport, Bolden, & Chung, 2011; Woodhead & Faulkner, 2008). For example, developmental expertise can be helpful in determining the appropriateness of methods for youth populations. However, evaluator characteristics, such as level of experience, highest degree obtained, and evaluation role (external/internal) are related to evaluation design preferences (Azzam, 2007; Christie, 2003). It is unclear to what extent evaluator characteristics and developmental expertise are related and subsequently relate to developmentally appropriate practices.

There are three components of the developmental context that are applicable for developmentally appropriate practice in evaluation. The first component of the developmental context involves knowledge of principles and theories of development (Berry et al., 2009). Culturally competent use of theories requires using culturally specific theory (AEA, 2011); in the case of youth programs, this means the use of developmental theories and principles. Principles of development explain how the environment, the individual, and the interaction between the environment and individual explain development over time. For example, Bronfenbrenner's bioecological systems theory (Bronfenbrenner & Morris, 2006) describes five different systems of contexts that affect the child directly and indirectly. Furthermore, his theory discusses proximal processes—the bidirectional interactions between the individual and an aspect of the environment—and how they are critical for human development (Bronfenbrenner & Morris, 2006). Knowledge of these principles and theories of development can help evaluators in designing and conducting evaluations, such as by helping identify appropriate outcomes and contexts to examine (Berry et al., 2009).

The second component of the developmental context involves domains of development, including cognitive, socioemotional, and physical domains (Berry et al., 2009). The cognitive domain refers to intellectual or mental development, such as thinking, memory, reasoning, problem-solving, language, and perception. The socioemotional domain refers to relationship skills, social awareness, self-management, self-awareness, and responsible decision-making (CASEL, 2012). The physical domain refers to development of body structure, including sensory/motor development and coordination of perception and movement. When evaluators understand the developmental stages of participants across these three domains, they are better equipped for designing and conducting developmentally appropriate evaluations. For example, knowledge of the cognitive stage of participants can help evaluators accommodate the reading level ability for construction of a paper survey. Knowledge of the socioemotional stage of participants can determine whether focus groups or interviews would be better suited for participants. Also, knowledge of the physical stage of participants can determine whether computer surveys, which require the use of fine motor skills for using a mouse and keyboard, are appropriate (Berry et al., 2009).

The third component of the developmental context is the age of participants (Berry et al., 2009). In general, age of participants can help explain the developmental stage of participants; for example, children under the age of seven typically are not at an adequate reading level for paper or online surveys. However, while age is typically used as a determinant of whether a data collection method is developmentally appropriate, the issue becomes complicated when considering children and adolescents from diverse populations (e.g., low-income, cultural and ethnic minorities, those with mental, emotional, or physical challenges; Shonkoff & Phillips, 2000). Disadvantaged youth may not be at the same developmental stages as their more advantaged counterparts. Age is a simple factor to consider when designing and conducting

evaluations but may not be sufficient to ensure a developmentally appropriate evaluation.

In sum, designing and conducting developmentally appropriate evaluations requires consideration of both age and developmental stages of participants, as well as knowledge and application of developmental theories and principles. These three components of the developmental context are interrelated and need to be considered in relation to all aspects of the program and evaluation context, especially for youth programs (Berry et al., 2009). The following describes how the developmental context affects three main components of the evaluation: the approach, the design, and the methods.

Evaluation Approach

When designing and conducting an evaluation, one choice an evaluator must make is the extent that stakeholders (i.e., individuals that have an interest in, or are affected by, the evaluation and/or its results) are involved. Evaluators typically include stakeholders for pragmatic reasons (i.e., to enhance problem-solving), political reasons (i.e., to enhance social justice issues), or epistemological reasons (i.e., to increase validity of knowledge generated). Stakeholders involved can include policy-makers and decision-makers, implementers and deliverers, target populations and intended beneficiaries, and civil society and citizens (Daigneault & Jacob, 2009).

One evaluation approach that involves stakeholders is participatory evaluation. In particular, many researchers and evaluators espouse that children and adolescents should be involved throughout the research process (Hart, 1992; Gong & Wright, 2007; Lerner, 2006; Powers & Tiffany, 2006; Sabo Flores, 2008). They argue that participatory methods can also reduce the harmful power imbalances between adult evaluators and child and adolescent participants (Christensen & James, 2008; Morrow & Richards, 1996), are more ethical for children and adolescents (Gallagher, 2008), and provide many benefits for children and

adolescents, including knowledge about the evaluation process and improvements in self-esteem, decision-making, and problem-solving skills (Kirby, 1999; Weithorn & Scherer, 1994).

Participatory research involving youth can range in levels from simply including participants as a data source (very low youth involvement) to youth giving input on the evaluation process (medium-low youth involvement) to youth leading the evaluation (highest youth involvement; Coad & Evans, 2007; Sabo Flores, 2008).

One way evaluators can ensure cultural competence is through inclusion of cultural expertise on the evaluation team (Chouinard & Cousins, 2009). While program implementers and deliverers have knowledge regarding the youth culture, the youth themselves (i.e., the target population and intended beneficiaries) will be most knowledgeable and have the most expertise of all stakeholder groups (Davis, 1998; Fine & Sandstrom, 1988). Thus, inclusion of youth in the evaluation process can enhance the evaluation team's cultural knowledge of the program and help ensure a developmentally appropriate evaluation.

Evaluation Design

Another choice evaluators have to make is the design of the evaluation. Examples of evaluation designs include descriptive (e.g., case study, observational), correlational (e.g., cohort study, cross-sectional study), quasi-experimental (e.g., nonequivalent control groups design, regression discontinuity design), experimental (i.e., experiment with true randomization), and meta-analysis designs (Crano & Brewer, 2002). There has been a debate within the field of evaluation on what constitutes credible evidence (Donaldson, Christie, & Mark, 2009), with some evaluators arguing for RCTs as the "gold standard" (Petrosino, 2003; Scott-Little, Hamann, & Jurs, 2002) and others questioning the superiority of the RCT (Jacobs, 2004; McCall & Green, 2004). In particular, evaluators of youth programs have argued that RCTs may not always fit the program or the program's need for evaluation (McCall & Green, 2004). Evaluations can go

beyond the study of outcomes to explore the needs of the program and the fidelity of implementation (McCall & Green, 2004; O'Donnell, 2008). Rather than arguing for a “gold standard” evaluation design, the evaluation design should be determined by the program, evaluation, and developmental contexts, as well as the evaluation questions the program has.

One of the components of the developmental context, principles and theories of development, discusses how development is the interaction between the person and their contexts over time (Berry et al., 2009). To measure and understand this development, longitudinal designs must be used (Bronfenbrenner & Morris, 2006; Elder, 1998; Lerner, 2006). As one example, Bronfenbrenner's bioecological systems theory requires the use of a process-person-context-time research design, indicating the need for longitudinal designs to examine development (Bronfenbrenner & Morris, 2006; Tudge, Mokrova, Hatfield, & Karnik, 2009). Furthermore, many child and adolescent programs have program models suggesting intermediary effects, as well as long-term effects after services have ended, both of which require longitudinal or multi-phase designs to formally assess (Shonkoff & Phillips, 2000). Thus, utilization of a longitudinal design in the context of youth programs can help ensure developmentally appropriate evaluations.

Evaluation Methods

Lastly, evaluators also need to make choices regarding data collection methods, measures, procedures, and data analysis. Developmentally appropriate practice in evaluation requires careful consideration of the methods used to ensure reliable data and valid inferences (Bell, 2007; Christensen & James, 2008; Fargas-Malet, McSherry, Larkin, & Robinson, 2010; Punch, 2002). For younger children, some data collection methods may be developmentally inappropriate; for example, interviews and focus groups are considered developmentally inappropriate for children under the age of 4 and paper and online surveys are considered

developmentally inappropriate for children under the age of 7. For older children, data collection methods may need to be adjusted to accommodate the age and developmental stages of participants; for example, surveys for children ages 7-12 require simple, direct, and personalized questions and only a few response options.

Regardless of the age of participants, it is recommended that evaluators conduct pre-testing of their interview, focus group, and survey instruments (Scott, 2008). This can include cognitive techniques, such as asking the child to ‘think aloud,’ retrospective probing, and convening a panel of researchers and other experts. Additional techniques include pilot testing the instruments and checking for inconsistencies in individuals’ responses (Bell, 2007). Furthermore, specifically for survey instruments, levels of item non-response, internal reliabilities of scales, unexpected results can reveal questions that youth have difficulty understanding (Bell, 2007). These practices help ensure children and adolescents understand the survey questions and are providing highly valid and reliable responses.

The following sections highlight three common traditional methods used in evaluations—surveys, interviews and focus groups, and observations and ethnography—and how these data collection methods can be modified to accommodate children and adolescents. As noted previously (see section on Developmental Context), age is typically used as the determinant of whether a data collection method is developmentally appropriate but may not be sufficient to ensure the data collection method is appropriate given the developmental stages of participants. This paper focus on age for sake of clarity, but evaluators should also take into consideration other factors that may influence developmental stages (e.g., disadvantaged youth; gifted and talented youth; youth with mental, emotional, or physical challenges). The ages given are for typically developing children and adolescents.

Surveys. There is general agreement that paper and online surveys are feasible with

children from the age of 7 onwards, with careful adaptation (Bell, 2007; Borgers, De Leeuw, & Hox, 2000; Scott, 2008). Surveys with children between the ages of 7 and 12 need more careful adaptation than surveys with adolescents between the ages of 13 and 18. Standardized adult questionnaires may be used with adolescents, with careful attention to question wording (De Leeuw, 2011). See Table 1 for developmentally appropriate practice in survey methodology. Specific modification suggestions are provided for children (ages 7-12) and adolescents (13-18) for question wording, response options, and general survey design.

Table 1.

*Developmental Appropriateness in Survey Methodology***Question Wording (de Leeuw, 2011)***General recommendations:*

- Instructions and questions should be written at a grade level at or below the reading grade level of participants (Borgers et al., 2000)
- Questions should be clear and unambiguous (i.e., questions should be interpreted the same by all participants; Bell, 2007; Ogan et al., 2013)
- Questions should not be double-barreled (i.e., include more than one question, use logical operators such as or/and/but)
- Complex questions (e.g., questions with semicolons, colons, brackets, parentheses, subordinate clauses) should be avoided
- Questions should not be negatively formulated (e.g., avoid double negatives)
- Vague quantifiers (e.g., sometimes, often) should be avoided

Recommendations for children (7-12):

- Questions need to be direct and personalized because children tend to be literal (Bell, 2007; Borgers et al., 2000)
- Questions should have neutral wordings to avoid suggestibility and social desirability (Bell, 2007; Borgers et al., 2000; Borgers et al., 2003; Scott, 2008)
- Questions that cannot be directly retrieved from memory should be avoided due to memory of children not fully grown (Bell, 2007; Ogan et al., 2013)
- Avoid questions that deal with a reference period in the past

Recommendations for adolescents (13-17):

- Indirect and depersonalized questions can be used (e.g., “Most people...”, “People my age...”)
- Suggestibility and social desirability are less of a concern, but suggestively phrased questions or statements should still be avoided (Borgers et al., 2000; Borgers et al., 2003)
- Questions that cannot be directly retrieved from memory can be used, but adequate time should be provided due to lower processing speeds compared to adults
- Questions with a defined reference period can be used; questions with an undefined reference period should be avoided

Response Options (de Leeuw, 2011)
General recommendations:

- Response options should be completely labeled (Bell, 2007; Borgers et al, 2003)
- Vague quantifiers (e.g., sometimes, often) should be avoided (Ogan et al., 2013)
- Neutral options should only be included if it can be clearly labeled
- “Don't Know” options should only be provided if there are clear indications (Bell, 2007)
- Visual response options (e.g., smileys) can be used (Bell, 2007)
- Response options should be completely balanced (i.e., equal number of positive and negative options)

Recommendations for children (7-12):

- No more than four response options should be provided (Bell, 2007)

Recommendations for adolescents (13-18):

- More than four response options can be provided (Bell, 2007)
-

Survey Design*General recommendations:*

- Use a clear and extensive introductory text to prevent item nonresponse (Borgers & Hox, 2001)
- Visual stimuli can help make the survey more concrete and interesting (Scott, 2008)
- Privacy should be ensured if surveyed as a group to avoid issues with sensitivity and group norms (de Leeuw, 2011)
- Survey length should be short to prevent motivation and boredom issues (Borgers et al., 2000; Borgers & Hox, 2001; Ogan et al., 2013)

Recommendations for children (7-12):

- Open-ended questions should be avoided due to low response rate (Ogan et al., 2013)

Recommendations for adolescents (13-18):

- Open-ended questions can be used (Ogan et al., 2013)
-

Note: This table was largely adapted from de Leeuw's (2011) report, particularly in terms of question wording and response options. Citations are included for specific recommendations when other researchers provided the same recommendation, or provided a recommendation that de Leeuw did not provide.

Interviews and Focus Groups. Interviews and focus groups can be used successfully with children from age four onwards with careful consideration (Eder & Fingerson, 2003; Larsson & Lamb, 2009). Developmentally appropriate practice in interviews and focus groups can be seen in Table 2. In particular, recommendations are given for whether to choose interviews or focus groups, general conduct of evaluators, question wording, and specific focus group methodology. Furthermore, power dynamics are particularly salient during interviews and focus groups which may interfere with collecting valid and reliable data from children and adolescents. Methods for reducing the power differential are also provided in Table 2.

Table 2.

Developmental Appropriateness in Interviews and Focus Group Methodology

Choosing Interviews or Focus Groups
<ul style="list-style-type: none"> • Interviews should be prioritized if the topic is sensitive or complex in nature (Lewis, 1992; Mauthner, 1997) • Youth may feel more comfortable in focus groups than individual interviews (Eder & Fingerson, 2003; Gallagher, 2009; Graue & Walsh, 1998; Lewis, 1992) • Participants in focus groups may become silent or talk over other participants, detracting from the benefits of using focus groups (Lewis, 1992) • Focus groups should be used when degree of consensus among group members is likely low (Clark, 2009) • Focus groups allow observation of dynamic interchange in group settings and understanding of shared meanings (Gallagher, 2009) • Younger children (4-7) may do better in individual interviews or in pairs or triads than larger focus groups (Einarsdóttir, 2007; Graue & Walsh, 1998)
General Conduct
<ul style="list-style-type: none"> • Interviews and focus groups should be conducted in contexts that are natural and familiar to youth, such as the playground or cafeteria (Eder & Fingerson, 2003; Gallagher, 2009; Parkinson, 2001; Scott, 2008) • Interviews and focus groups may further be aided by being conducted within a larger activity with which respondents are already familiar (e.g., recess, show and tell, circle time) or by creating games similar to ones children are already familiar with (Eder & Fingerson, 2003; Graue & Walsh, 1998; Parkinson, 2001) • The interviewer should establish trust with participants prior to the interview (Eder & Fingerson, 2003; Fargas-Malet et al., 2010; Parkinson, 2001) • Length of interviews or focus groups should be maximum 45 minutes for children (4-12) and 60 minutes for young adolescents (13-16; Lewis, 1992) • Props may be helpful but should be pilot tested to ensure they do not distract or decrease accuracy of responses (Fargas-Malet et al., 2010; Graue & Walsh, 1998; Larsson & Lamb, 2009; Mac Naughton et al., 2001; Mauthner, 1997; Parkinson, 2001) • Visual stimuli may be helpful in making questions more concrete than verbal representation alone, provide a visual aid in response options when the question is close-ended, and aid in memory recall or recognition (Larsson & Lamb, 2009; Mac Naughton et al., 2001; Mauthner, 1997; Parkinson, 2001; Scott, 2008)
Question Wording
<ul style="list-style-type: none"> • Semi-structured protocols should be prioritized to make the interview or focus group feel more natural for youth, particularly for younger children (4-7; Eder & Fingerson, 2003; Einarsdóttir, 2007; Fargas-Malet et al., 2010; Graue & Walsh, 1998; Parkinson, 2001) • Younger children (4-7) should be probed often to ensure they understand the question as they will provide responses even if they do not know what the question or their answer means (Einarsdóttir, 2007) • <i>Note: All items regarding question wording for surveys (see Table 1) also apply to interviews and focus groups</i>
Methods for Reducing the Power Differential
<ul style="list-style-type: none"> • Using children as interviewers (i.e., peer interviewing; Eder & Fingerson, 2003; Lewis, 1992)

- Using focus groups because children are more comfortable with peers (Eder & Fingerson, 2003)
- Employing reciprocity (i.e., self-disclosing; Eder & Fingerson, 2003)
- Empowering children and adolescents through the interviews or focus groups (Eder & Fingerson, 2003)
- Collectively negotiating interpretation of results (Eder & Fingerson, 2003)

Focus Group Methodology

- Groups should be around 4-6 participants, on the lower end for younger children (Einarsdóttir, 2007; Gibson, 2007; Lewis, 1992; Parkinson, 2001; Scott, 2008)
 - Group should be comprised of same-sex participants (Gibson, 2007; Lewis, 1992; Mauthner, 1997; Scott, 2008)
 - Groups should be comprised of participants no more than 2 years apart in age (Gibson, 2007; Lewis, 1992; Mauthner, 1997; Scott, 2008)
 - Using breaks (e.g., food or water break) can help keep youth focused and provide extra non-structured informal time with you to improve trust (Fargas-Malet et al., 2010)
-

Observations and Ethnography. In the past, methods for gathering data about children and adolescents were limited to observations and ethnography (Punch, 2002). Reasons for limiting data collection methods to observations and ethnography included: adults were credited with greater knowledge, experience, and power; interviewing minors poses practical and ethical issues which researchers might want to avoid; children were commonly believed to lack the developmental abilities to provide reliable and valid data (Punch, 2002; Scott, 2008).

Researchers and evaluators no longer believe these to be true and thus the field has shifted from collecting data solely *on* children and adolescents (i.e., ethnography and observation, proxy reports) to also collecting data *from* children and adolescents (i.e., collecting data directly from children and adolescents).

However, while knowledge about children may be best obtained from children (Gallagher, 2009; Hill, 1997; Scott, 2008), observations and ethnography are still valuable methods to use, particularly for children under the age of four who cannot participate in surveys, interviews, or focus groups (Tisdall, Davis, & Gallagher, 2009). In fact, observations and ethnography are considered less disruptive of children and adolescents' lives. Participant

observation, where the adult evaluator participates in the child or adolescent's life while conducting observations, is particularly popular as they can more directly explore what children say and do. Participant observation can only be conducted from pre-school age (3-6) onwards (Fine & Sandstrom, 1988; Tisdall et al., 2009).

Present Study

Accommodating differences between children, adolescents, and adults requires a culturally competent approach to ensure the evaluation of youth programs are developmentally appropriate. One strategy for developmentally appropriate practice is to include developmental expertise onto the evaluation team. Evaluators with developmental expertise can better ensure that evaluations of youth programs are aligned with the developmental context, such as by ensuring methods used are developmentally appropriate given the age and developmental stages of program participants.

However, it is not yet clear whether and how evaluators of youth programs are adapting their evaluation designs according to the developmental context. A previous content analysis of youth-focused evaluations suggests that evaluators may not be employing practices aligned with the developmental context (Bialosiewicz, 2013). Thus, this study aims to answer the following research questions:

1. Do evaluators adapt their evaluations according to the developmental context?
2. Do evaluators with more developmental expertise adapt their evaluations better for young populations?
3. Do other evaluator background characteristics relate to evaluators adapting their evaluations to the developmental context?

To answer these questions, practicing evaluators participated in an experimental simulation study where they responded to a Request for Proposal (RFP) of a hypothetical tutoring program which

differs on the age of participants (i.e., child, adolescent, college-aged adults).

Methods

Participants

Participants in the study consisted of two distinct samples of evaluators from the American Evaluation Association (AEA) directory. The first sample was derived from the entire list of members in the PreK-12 Education and Youth Focused Evaluation (YFE) Topical Interest Groups (TIG) and will subsequently be referred to as Youth TIG evaluators. There a total of 603 members in the PreK-12 TIG, 206 in the YFE TIG, and 68 in both the PreK-12 and YFE TIGs for a total sample of 877 Youth TIG evaluators. The second sample was randomly selected from a list of 1,000 evaluators and will subsequently be referred to as Other TIG evaluators. There were 119 Youth TIG evaluators in the Other TIG evaluators group, resulting in a total of 881 Other TIG evaluators.

Of the 1,758 evaluators in the sample, 475 (27%) agreed to participate. The three (0.6%) individuals who did not agree to participate had differing reasons for not participating: one did not conduct evaluations (PreK-12 TIG), one did not feel comfortable responding to a scenario involving education or students (Other TIG), and one did not say why they did not wish to participate (Both PreK-12 and YFE TIGs). Youth TIG evaluators ($n = 271$, 31% response rate) were more likely to participate in the survey than Other TIG evaluators ($n = 204$, 23% response rate), $\chi^2(1) = 11.94$, $p < .001$.

Of the 475 evaluators who agreed to participate, 279 (58.7%) finished the entire survey and it took them a median of 27 minutes to complete the survey. Youth TIG evaluators (60.9%) were more likely to finish the entire survey than Other TIG evaluators (53.4%, $\chi^2[1] = 3.96$, $p = .047$). The evaluators who did not finish the survey were retained in analyses, which accounts for the differing sample sizes throughout the results section. Furthermore, evaluators were evenly

randomly assigned to the three survey conditions and there were no differences in whether evaluators finished the survey by randomization ($\chi^2[2] = 2.93, p = .231$).

Descriptive characteristics of the evaluation sample can be found in Table 3 and developmental expertise of the participants can be found in Table 4. Youth TIG evaluators reported practicing evaluation for more years ($M = 3.79, SD = 1.13$) than Other TIG evaluators ($M = 3.41, SD = 1.24, t [291] = 2.70, p = .007$), were more likely to have a PhD (64%) than Other TIG evaluators (50%, $\chi^2 [1] = 5.06, p = .024$), and were more likely to be an external evaluator (57%) than Other TIG evaluators (40%, $\chi^2 [2] = 7.71, p = .021$). The three categories of developmental expertise (i.e., number of evaluations, number of classes, knowledgeable in development) were combined at the child, adolescent, and adult levels for easier comparison. Youth TIG evaluators had higher levels of child expertise ($M = 2.82, SD = 1.09$) than Other TIG evaluators ($M = 1.99, SD = .91, t [291] = 6.80, p < .001$) and higher levels of adolescent expertise ($M = 3.01, SD = 1.04$) than Other TIG members ($M = 2.19, SD = .94, t [291] = 6.83, p < .001$); there were no differences in adult expertise across TIGs. Furthermore, there were no differences in evaluator background characteristics or developmental expertise across randomization conditions.

Table 3.

Evaluator Experience Characteristics (n = 293)

Years Conducting Evaluations	
Less than 1 year	0.7%
1 to 5 years	22.8%
6 to 10 years	22.4%
11 to 15 years	20.1%
More than 15 years	34.0%
Evaluator Role	
Primarily internal	22.4%
Primarily external	49.3%
A mix of both	25.9%
Other	2.4%
Highest Degree	
Bachelor's	2.4%

Master's	40.1%
Doctorate	57.5%
Degree Type (check-all-that-apply)	
Education	41.8%
Psychology and Development	28.2%
Health	20.8%
Evaluation	18.1%
Sociology	17.7%
Policy and Political Science	16.1%
Statistics and Mathematics	13.4%
Organization and Business	8.1%
Economics	3.7%
Anthropology	2.7%
Other	8.5%

Table 4.

Developmental Knowledge and Experience (n = 282)

	0 (%)	1-3 (%)	4-6 (%)	7-9 (%)	10+ (%)	M (SD)
Number of Evaluations						
Children	34.4	27.4	13.2	6.6	18.4	2.47 (1.48)
Adolescents	24.2	29.0	15.4	8.9	22.5	2.76 (1.48)
Young Adults	30.3	26.4	15.5	10.2	17.6	2.58 (1.48)
Tutoring Programs	50.2	33.2	7.8	3.5	5.3	1.81 (1.08)
Number of Classes or Professional Development Workshops						
Children	36.6	34.9	11.3	6.2	11.0	2.20 (1.30)
Adolescents	33.9	38.4	9.2	5.1	13.4	2.26 (1.33)
Young Adults	41.4	34.5	10.3	2.8	11.0	2.08 (1.27)
	Not at all			Extremely		
How knowledgeable are you in developmental principles/theories?						
Child Knowledge	14.4	24.3	32.5	20.2	8.6	2.84 (1.16)
Adolescent Knowledge	12.0	18.5	36.6	18.5	14.4	3.05 (1.20)
Young Adult Knowledge	12.4	21.4	36.2	18.6	11.4	2.95 (1.16)

Measures and Procedures

Figure 1 shows the procedures of the simulation study. The following details the specific measures used.

Figure 1. Study Procedures.

Hypothetical program description. Evaluators were given a program description of a hypothetical tutoring program. A tutoring program was chosen because it was a program that could have a range of developmental outcomes examined (e.g., cognitive, socioemotional) and it could rationally be provided to participants of all ages (e.g., children and adolescents in K-12 and adults in college). Evaluators were randomly assigned to one of three participant age groups (i.e., children aged 4 to 12 in grade levels K-6, adolescents aged 13 to 17 in grade levels 7-12, or young adults aged 18 to 26 in college). The program description, as well as the budget for a three-year evaluation, was based on a content analysis of evaluation reports, request for proposals (RFPs), and other related documents to ensure the hypothetical program descriptions

mimic real-world tutoring programs. Evaluators were told they had been hired to evaluate both the implementation and impact of the program in a three-year evaluation with a total budget of \$200,000. They then responded to questions regarding the evaluation approach, design, and data collection methods they were likely use for the evaluation.

Evaluation approach. Evaluators rated how uninvolved or involved various stakeholders would be in the design of the evaluation. Stakeholders included school administration, teachers, parents, program staff, tutors, tutees, program designers, district personnel, funders, developmental consultants, math consultants, and an option for “other” stakeholders; all options were randomly presented to evaluators, with the “other” option as the last option.

Evaluation designs. Evaluators then chose their first and second choices of evaluation designs. Evaluation designs options included experimental, quasi-experimental, descriptive/correlational, ethnographic, case study, and an option for “other” designs where evaluators could write-in a different evaluation design; all options were randomly presented to evaluators, with the “other” option as the last option. Evaluators also responded to which type of process, outcome/impact, and cost-effectiveness questions they would try to answer in their evaluation. These questions were derived primarily from Better Evaluation (2016).

Evaluation methods. Evaluators were provided a matrix of evaluation data collection methods (i.e., surveys, interviews, focus groups, observations, case studies, and an option for “other” methods where evaluators could write-in a different evaluation method) and data collection sources (i.e., tutees, tutors, program staff, parents, teachers, school administrators) to choose from; all data collection methods were randomly presented to evaluators, with the “other” option as the last option. Archival data and case studies were also included for tutees and tutors only. The main purpose of these questions was to determine whether evaluators would collect data directly from program participants and to then ask specific questions regarding data

collection tools, procedures, analysis, interpretation, and dissemination of data.

If evaluators indicated they would either issue surveys or conduct interviews/focus groups with the tutees of the program, they then responded to questions about their survey practices or interview/focus group practices. However, if they indicated they would do both surveys and interviews/focus groups, evaluators were randomly assigned to answer questions about only one. Across both conditions, they were asked (a) what would be assessed (e.g., implementation, satisfaction, math knowledge and skills), (b) whether multiple surveys or interview protocols would be created for different age groups, (c) what techniques would be used to design the survey or protocols (e.g., previously validated scales/protocols, creation of new scales/protocols), (d) what stakeholders would be included when designing or interpreting the surveys or protocols, (e) which tutees the surveys or protocols would be administered to, and (f) what techniques would be used prior to data collection (e.g., pilot testing, cognitive interviews, think-alouds, panels of experts).

If evaluators indicated they would not collect data directly from participants, they were asked an open-ended question on their rationale for not doing so.

Surveying program participants. For the survey condition, additional questions that were asked included: (a) how many survey measurements would be conducted across the three-year evaluation (i.e., baseline or post-test only, baseline plus one follow-up, baseline plus multiple follow-ups), (b) the average number of response options used, (c) how the survey would be administered, and (d) what data analytic techniques they would use.

Conducting interviews and/or focus groups with participants. For the interview/focus group condition, additional questions asked included: (a) how many sets of interviews or focus groups they would conduct across the three-year evaluation (i.e., one set or multiple sets), (b) the structure of the interviews or focus groups (i.e., unstructured, semi-structured, structured), (c)

who would conduct the interviews or focus groups, (d) how long the interviews or focus groups would last, (e) how many tutees would be in the focus groups, and (f) whether focus groups would be differentiated by age, sex, and/or race/ethnicity.

Developmental appropriate questions. At the end of the survey, evaluators were asked to respond to three open-ended questions about their general evaluation practices. These questions asked how the age of participants in the program affect their general evaluation practices when (a) planning, (b) implementing, and (c) reporting and ensuring use of the evaluation.

Evaluator background questions. To determine the extent to which evaluator background characteristics and developmental expertise are related, and subsequently affect evaluation design preferences, evaluators responded to a series of background questions at the end of the survey. Evaluation background characteristics included their level of experience, highest degree obtained, and evaluation role (external/internal). Furthermore, evaluators were asked what level of experience they have with conducting evaluations of youth programs. Developmental expertise was determined by asking about the number of evaluations they had conducted across the three age groups (i.e., child, adolescent, young adult), the number of classes or professional developmental workshops that involved each of the three age groups, and how knowledgeable they were in developmental principles and theories across the three age groups and averaged into a composite score.

Findings

The following answers the three research questions and details the primary findings of the evaluation practices evaluators used across the child, adolescent, and adult conditions.

1. Do evaluators adapt their evaluations according to the developmental context?

Evaluation Approach: Evaluators were less likely to involve program beneficiaries

when they are children or adolescents than when they are adults. When responding to questions regarding the levels of involvement there would be across a wide range of stakeholders, evaluators in the child and adolescent conditions were less likely to involve program beneficiaries (i.e., tutees and tutors) than evaluators in the adult condition (see Figure 2). Evaluators were less likely to involve tutees in the child condition ($M = 2.50$, $SD = 1.19$) than the adolescent condition ($M = 2.73$, $SD = 1.03$), and evaluators in both the child and adolescent conditions were less likely to involve tutees than evaluators in the adult condition ($M = 2.96$, $SD = 1.04$), $F(2, 333) = 5.10$, $p = .007$. Furthermore, evaluators were less likely to involve tutors in the child condition ($M = 2.81$, $SD = 1.16$) than evaluators in the adult condition ($M = 3.26$, $SD = .96$), $F(2, 333) = 5.63$, $p = .004$. There were no differences across conditions for any other stakeholders, including the developmental and math consultants.

Figure 2. Levels of tutee and tutor involvement across conditions ($n = 336$)

Evaluation Design: Evaluators did not change their evaluation designs based on the age of participants. Evaluators were asked to rank in order which evaluation design they would use in their evaluation of the hypothetical program. There were no differences between evaluators in the child, adolescent, or adult conditions on the design choices they would use. Most evaluators ranked quasi-experimental design first (36.7%) followed by

descriptive/correlational (22.7%) and experimental (15.3%). Few evaluators chose a case study (8.3%) or ethnographic (4.2%) design. However, 12.8% of evaluators chose to write in another design, with the majority indicating they would use a mixed methods design (36.8%), a collaborative/empowerment/participatory design (5.9%), a developmental design (5.9%), a qualitative design (5.9%), or a utilization-focused design (5.9%).

Evaluation Methods: Evaluators are less likely to collect data directly from participants when they are children than when they are adults. Evaluators were asked to choose what methods they would use for each participating stakeholder. In general, evaluators in the child condition were less likely to directly collect data (i.e., survey, interview, focus groups) and more likely to indirectly collect data (i.e., observations, case study, archival data) from tutees than evaluators in the adult condition (see Figure 3). Specifically, evaluators in the child condition were less likely to survey tutees (64.0%) than evaluators in the adolescent (92.3%) or adult (96.0%) conditions, $\chi^2 (2) = 42.50, p < .001$. Evaluators in the child condition were also less likely to conduct focus groups with tutees (61.8%) than evaluators in the adolescent (82.9%) or adult (90.9%) conditions, $\chi^2 (2) = 23.59, p < .001$. On the other hand, evaluators in the child condition were more likely to conduct observations (95.8%) than evaluators in the adolescent (88.5%) or adult (75.6%) conditions, $\chi^2 (2) = 16.10, p < .001$.

Interestingly, evaluators in the child (93.0%) and adolescent (91.2%) conditions were more likely to survey tutors than evaluators in the adult condition (79.8%), $\chi^2 (2) = 9.02, p = .011$. Evaluators in the child (91.0%) or adolescent (87.8%) conditions were also more likely to conduct focus groups with tutors than evaluators in the adult condition (67.7%), $\chi^2 (2) = 8.42, p = .015$. Furthermore, evaluators in the child condition were more likely to collect tutors' archival data (69.5%) than evaluators in the adolescent (40.0%) or adult (47.1%) conditions, $\chi^2 (2) = 19.37, p < .001$.

Figure 3. Percent of method use for tutors and tutees across condition ($n = 304$)

Methods: Evaluators differed across conditions in the administration techniques

they would use. Evaluators were asked to indicate which of the following survey administration techniques they would use: observe the survey administration, administer the survey orally, have program staff administer survey, or use an online survey. Evaluators in the child condition were more likely to have program staff administer the survey ($\chi^2[2] = 9.40, p = .009$) and use oral administration ($\chi^2[2] = 30.28, p < .001$) and less likely to use online administration ($\chi^2[2] = 15.54, p < .001$) than evaluators in the adult condition.

Evaluators were also asked who would conduct the interviews: the evaluation team (86.5%), interview specialists (4.5%), program staff (1.3%), tutors (1.3%), or tutees (0.6%). Evaluators in the child condition were less likely to use the evaluation team (75.5%) than evaluators in the adolescent (93.9%) or adult (90.6%) conditions and more likely to use interview specialists (11.3%) than evaluators in the adolescent (0.0%) or adult (1.9%) conditions, $\chi^2(2) = 8.53, p = .014$ (see Figure 4).

Figure 4. Percent of survey administration use across condition ($n = 124$)

Interviews: Evaluators were more likely to have shorter interviews, fewer participants in focus groups, and focus groups comprised of participants of similar ages when participants were children than when they were adolescents or adults. Evaluators indicated how long the interview would be, from less than 15 minutes (3.9%), 15-30 minutes (27.0%), 30-45 minutes (30.9%), 45-60 minutes (27.6%), 60-75 minutes (7.9%), or more than 75 minutes (2.6%). Evaluators in the child condition indicated shorter interviews ($M = 2.60, SD = .89$) than evaluators in the adolescent ($M = 3.44, SD = 1.05$) and adult ($M = 3.50, SD = 1.21$) conditions, $F(2, 149) = 11.67, p < .001$.

Evaluators also indicated how many people would be in the focus groups, from 2-3 students (1.7%), 4-5 students (22.5%), 6-7 students (40.8%), 8-9 students (24.2%), 10-11 students (8.3%), or 12 or more students (2.5%). Evaluators in the child condition indicated fewer

participants ($M = 2.63$, $SD = .85$) than evaluators in the adolescent ($M = 3.44$, $SD = .94$) and adult ($M = 3.56$, $SD = 1.03$) conditions, $F(2, 117) = 11.08$, $p < .001$.

Last, evaluators indicated whether the focus groups would be comprised of tutees of the same sex (15.3%), same age (70.3%), same race (11.9%), or if their groups would be undifferentiated by any demographic or construct (27.1%). Evaluators in the child (84.2%) and adolescent (77.5%) conditions were more likely to create focus groups of the same age than evaluators in the adult condition (50%), $\chi^2(2) = 12.42$, $p = .002$.

Surveys and Interviews: There were no differences across conditions for any other survey- or interview-specific methodology practices. There were no differences across conditions for any of the other survey specific methodology questions asked beyond the administration technique used. First, evaluators indicated whether they would use a baseline or post-test survey only (3.2%), a baseline survey and one-follow-up survey (28.0%), or a baseline survey and multiple follow-up surveys (68.8%); however, there may be no differences due to the simulation study involving a three-year evaluation. Second, evaluators indicated whether they would use 2-4 (35.8%), 5-7 (62.6%), 8-10 (1.63%), or 11+ (0%) response options in the survey; there may have been no differences due to a floor effect of response options typically used. Last, evaluators indicated whether they would test for internal consistency (62.1%), test-retest reliability (41.9%), face validity (70.2%), criterion validity (31.5%), construct validity (34.7%), use factor analysis techniques (51.6%), or test for moderators (34.7%).

There were also no differences across conditions for any of the other interview specific methodology questions asked beyond the length of interviews, number of participants in focus groups, or composition of focus groups. First, evaluators indicated whether they would conduct one set of interviews or focus groups (7.0%) or multiple sets of interviews or focus groups (93.0%); like the survey condition, there may be no differences due to the simulation study

involving a three-year evaluation. Second, evaluators indicated whether the interviews would be unstructured (0%), semi-structured (91.6%), or structured (8.4%); majority of evaluators seem to prefer semi-structured interviews regardless of the age of participants.

2. Do evaluators with more developmental expertise adapt their evaluations better for young populations?

Developmental expertise was operationalized in two different ways: membership in either the Youth-Focused Evaluation TIG or PreK-12 Education TIG (Youth TIGs) or levels of self-reported child experience (i.e., an aggregate of the number of evaluations with child participants conducted, the number of professional development courses taken about child development, and general knowledge of child development). Overall, there were few differences in evaluation practices across developmental expertise except for the differences discussed below. Furthermore, there were no significant interactions between condition and developmental expertise on any outcomes.

Youth TIG evaluators and evaluators with higher levels of child experience were less likely to involve tutors and tutees and more likely to observe tutees and tutors. Opposite from the expected findings, Youth TIG evaluators were less likely to involve tutees ($M = 2.57$, $SD = 1.12$) in the evaluation than Other TIG evaluators ($M = 2.96$, $SD = 1.04$), $t(334) = 3.19$, $p = .002$. Youth TIG evaluators were also less likely to involve tutors ($M = 2.91$, $SD = 1.09$) in the evaluation than Other TIG evaluators ($M = 3.27$, $SD = .96$), $t(334) = 3.13$, $p = .002$. Youth TIG evaluators were also less likely than Other TIG evaluators to involve tutees when designing protocols ($\chi^2(1) = 4.20$, $p = .041$) and to involve tutors when interpreting results ($\chi^2(1) = 6.94$, $p = .008$).

However, when examining within Youth TIG evaluators, YFE TIG evaluators were more likely to include tutees ($M = 3.20$, $SD = 1.02$) than Other TIG evaluators ($M = 2.96$, $SD = 1.04$)

and PreK-12 TIG Evaluators ($M = 2.35$, $SD = 1.09$), $F(3, 332) = 10.89$, $p < .001$. Furthermore, YFE TIG evaluators were more likely to include tutors ($M = M = 3.41$, $SD = .87$) than Other TIG evaluators ($M = 3.27$, $SD = .96$) and PreK-12 TIG Evaluators ($M = 2.73$, $SD = 1.13$), $F(3, 332) = 8.51$, $p < .001$.

Furthermore, contrary to expected findings, the amount of child experience evaluators had was negatively related to involvement of tutors ($r[291] = -.12$, $p = .035$) and positively related to involvement of math consultants ($r[290] = .12$, $p = .026$). Youth TIG members were also more likely to observe tutees (90.6%) than Other TIG members (81.1%), $\chi^2(1) = 4.04$, $p = .044$.

Evaluators with higher levels of child experience asked more outcome questions and conducted more survey follow-ups than evaluators with lower levels of child experience.

The only other differences in developmental experience were related to levels of child experience and the number of outcome questions asked and the number of survey measurements collected. Evaluators with greater levels of child developmental experience asked more outcome questions than evaluators with lower levels of child developmental experience ($r[291] = .14$, $p < .015$). Furthermore, evaluators who chose a baseline and multiple follow-up surveys had higher child experience ($M = 2.77$, $SD = 1.18$) than evaluators who chose a baseline survey and only one follow-up survey ($M = 2.17$, $SD = 1.03$), $F(2, 118) = 3.42$, $p = .036$.

3. Do other evaluator background characteristics relate to evaluation practices?

The practices of evaluators with a Master's degree differed in some respects to the practices of evaluators with a Doctoral degree. Evaluators with an MA were more likely to involve tutees ($M = 2.86$, $SD = 1.08$) than evaluators with a PhD ($M = 2.57$, $SD = 1.12$), $t(284) = 2.21$, $p = .028$. Evaluators with an MA were also more likely to involve a developmental consultant when designing protocols ($\chi^2[1] = 4.21$, $p = .041$) and interpreting results ($\chi^2[1] =$

4.47, $p = .034$). Evaluators with an MA were more likely to agree they would create different protocols for younger and older tutees (58.5%) than evaluators with a PhD (43.4%), $\chi^2 (1) = 4.51$, $p = .034$. Furthermore, evaluators with a PhD were more likely to test for internal consistency (69.4%) than evaluators with an MA (48.9%, $\chi^2 [1] = 4.11$, $p = .043$) and that evaluators who chose to analyze test-retest reliability had fewer years as an evaluator ($M = 3.10$, $SD = 1.27$) than evaluators who did not ($M = 3.79$, $SD = 1.12$; $t [120] = 3.18$, $p = .002$). Lastly, when examining the interaction between condition and education, evaluators in the child condition with a PhD were more likely to interview tutees (78.6%) than evaluators with an MA (54.8%), $F_{Condition*Degree} (1,148) = 4.40$, $p = .038$.

There were few differences between internal and external evaluators. External evaluators were more likely to involve tutors ($M = 3.27$, $SD = .97$) than internal evaluators ($M = 2.85$, $SD = 1.09$), $F (2, 283) = 4.15$, $p = .017$. External evaluators were also more likely to involve tutees ($M = 2.97$, $SD = 1.08$) than internal evaluators ($M = 2.51$, $SD = 1.13$), $F (2, 283) = 4.40$, $p = .013$. Furthermore, internal evaluators were more likely to not use the evaluation team (29.0%) than external evaluators (9.3%) or evaluators who reported both internal and external roles (11.6%), $\chi^2 (2) = 7.33$, $p = .026$.

There were no differences in evaluation practices across years of evaluation experience. Evaluators indicated the number of years they have been conducting evaluations (i.e., less than 1 year, 1 to 5 years, 6 to 10 years, 11 to 15 years, or more than 15 years). There were no differences in any evaluation practices across years of evaluation experience.

Many evaluators self-reported that the age of participants affects their general evaluation practices when planning, implementing, and reporting and ensuring use of evaluations.

At the end of the survey, evaluators were asked to respond to three open-ended questions about their general evaluation practices. These questions asked how the age of participants in the

program affect their general evaluation practices when (a) planning, (b) implementing, and (c) reporting and ensuring use of the evaluation. A total of 249 evaluators had a valid response for at least one of these open-ended questions and were subsequently analyzed. There were no differences between conditions or evaluator characteristics in any of the themes except for one aspect of ethics discussed below. The themes can be found in Table 6.

Table 6.

Themes of How Age Affects Evaluation Practice Among Evaluators (n = 249)

Theme	n (%)
Approach – Stakeholder Involvement	47 (18.9%)
Involve participants (youth)	24 (9.6%)
Involve parents	9 (3.6%)
Involve participants (elderly)	9 (3.6%)
Involve specialists (math or development)	3 (1.2%)
Design	22 (8.8%)
Reduced frequency and timing of data collection	14 (5.6%)
Increased flexibility in design and data collection	6 (2.4%)
Use designs that account for maturation	2 (0.8%)
Methods Used with Individuals Under a Certain Age	126 (50.6%)
Would affect the choice of method in general	58 (23.3%)
Would affect the administration technique used	40 (16.1%)
Would not survey individuals under certain age	20 (8.0%)
Would not interview individuals under certain age	10 (4.0%)
Would use creative methods (e.g., drawing, props) with individuals under certain age	8 (3.2%)
Would only observe or use other indirect methods with individuals under certain age	7 (2.8%)
Cognitive and Linguistic Capacities Differ Across Age	102 (41.0%)
Affect wording of questions	66 (26.5%)
Affects length of data collection protocols or surveys	34 (13.7%)
Attention spans as a key factor	17 (6.8%)
Maturity as a key factor	6 (2.4%)
General Data Collection Techniques	40 (16.1%)
Who collects the data (e.g., evaluator, staff, peers)	12 (4.8%)
Need to build trust and rapport with younger participants	7 (2.8%)
Social desirability is a concern with younger participants	7 (2.8%)
Use and types of incentives	7 (2.8%)
Pre-testing prior to data collection	4 (1.6%)
Use of warm-up activities	4 (1.6%)
Location of data collection	3 (1.2%)
Ethics	109 (44.8%)

Informed consent from parents	50 (20.1%)
Maintaining confidentiality and/or anonymity	36 (14.5%)
Informed consent or assent from minors	31 (12.4%)
IRB, FERPA, or other ethical guidelines	21 (8.4%)
Ensuring appropriate conduct with minors	8 (3.2%)
Recognizing power dynamics between adult evaluators and youth participants	8 (3.2%)
Reporting	137 (55.0%)
Tailored dissemination by age of participants	69 (27.7%)
Test for moderators (e.g., age, gender)	36 (14.5%)
Data visualization, interactivity, and technology important for younger participants	18 (7.2%)
Discussing limitations when using younger participants	8 (3.2%)
Using evaluation reporting practices to amplify the voices of minors	7 (2.8%)
<i>Age does not affect reporting practices*</i>	34 (13.7%)
<i>Unsure how age would affect reporting practices*</i>	5 (2.0%)
<i>* Not included in the sum of the Reporting theme</i>	

The most prevalent theme mentioned by evaluators was a general reporting theme in response to the third open-ended question asked. A total of 137 (55.0%) evaluators responded to this question, many of whom discussed how they would tailor dissemination by age of participants ($n = 69$, 27.7%), test for moderators such as age or gender ($n = 36$, 14.5%), or incorporate data visualization, interactivity, or technology for younger participants ($n = 18$, 7.2%). However, some evaluators ($n = 5$, 2.0%) were unsure how age would affect reporting practices and 34 (13.7%) evaluators stated age does not affect reporting practices. Many of the evaluators who stated age does not affect their reporting practices stated that program participants were not the main audience for the evaluation. For instance, one evaluator stated, “In evaluations with kids, I have never given them the results of the evaluation, or tried to ensure that they were using the evaluation, even in situations with high school kids.” Another evaluator stated, “I do not assume that young children will be interested in the results of an evaluation.”

The second most prevalent theme mentioned was discussion about what methods they would use with individuals under a certain age ($n = 126$, 50.6%). Specifically, many evaluators

($n = 58$, 23.3%) mentioned it would affect their choice of methods in general but did not specify how. Some evaluators ($n = 20$, 8.0%) discussed they would not survey individuals under a specific age, though the age of participants who would not be surveyed ranged widely (i.e., grade 3, grade 4, grade 5, elementary, grade 8, eight years of age, 12 years of age, 18 years of age). Ten (4.0%) evaluators discussed they would not use a focus group or interview under a certain age, though they were equally likely to not use both focus groups or interviews, not use only interviews, or not use only focus groups with younger children. Furthermore, seven (2.8%) evaluators discussed they would only use observations or other indirect measurements with young children and eight (3.2%) evaluators discussed they would try to use creative methods, like drawing or props, to elicit information from younger children.

The third most prevalent theme mentioned was ethics ($n = 109$, 44.8%). Many evaluators discussed how they needed to ensure informed consent from parents ($n = 50$, 20.1%) or informed consent or assent from minors ($n = 31$, 12.4%). Evaluators also discussed the importance of maintaining confidentiality and/or anonymity ($n = 36$, 14.5%) and following IRB, FERPA, and other ethical guidelines ($n = 21$, 8.4%). Furthermore, some evaluators discussed the importance of ensuring appropriate conduct with minors ($n = 8$, 3.2%) or recognizing the power differential between evaluators and youth ($n = 8$, 3.2%). Interestingly, no internal evaluators discussed IRB, FERPA, or other ethical guidelines whereas 10.1% of external evaluators and 9.5% of evaluators who state they are a mix of both internal and external evaluators discussed IRB, FERPA, or other ethical guidelines, $\chi^2(3) = 10.49$, $p = .015$.

Another prominent theme mentioned was how cognitive and linguistic capacities differed across ages ($n = 102$, 41.0%). Many mentioned how these capacities affected the wording of questions ($n = 66$, 26.5%) or the length of data collection protocols or surveys ($n = 34$, 13.7%). Some also discussed how attention spans ($n = 17$, 6.8%) or maturity levels ($n = 6$, 2.4%) affected

evaluations.

Themes less often mentioned were the evaluation approach ($n = 47$, 18.9%), design ($n = 22$, 8.8%), and general data collection techniques ($n = 40$, 16.1%). Some stakeholders discussed involving youth participants ($n = 24$, 9.6%), parents ($n = 9$, 3.6%), or math or developmental specialists ($n = 3$, 1.2%). Interestingly, some evaluators ($n = 9$, 3.6%) also discussed how elderly individuals may also affect their planning, implementation, or reporting of evaluations. While no evaluators discussed specific evaluation designs, some discussed the timeline of the evaluation and the need for reduced frequency and timing of data collection ($n = 14$, 5.6%), as well as increased flexibility in the design and in data collection ($n = 6$, 2.4%). Only two evaluators (0.8%) discussed that they would use designs that account for maturation. Furthermore, some evaluators discussed general data collection techniques such as who collects the data ($n = 12$, 4.8%), the need to build trust and rapport with younger participants ($n = 7$, 2.8%), and pre-testing ($n = 4$, 1.6%).

Discussion

The purpose of this study was to examine how evaluator practices differ across programs that different age groups. By asking evaluators to respond to questions about how they would evaluate a program where the only difference is whether the program serves children, adolescents, or adults, we can see the extent to which evaluators conduct developmentally appropriate evaluations in the context of youth programs.

The primary finding of this study is that evaluators are less likely to be participatory—both in the overall evaluation process and in data collection methods—when the program beneficiaries are children than when they are adults. Evaluators randomly assigned to the child condition were (a) less likely to involve tutees and tutors in the evaluation process, (b) less likely to use surveys with tutees, (c) less likely to use focus groups with tutors and tutees, (d) more

likely to observe tutees, and (e) more likely to collect archival data from tutors. Furthermore, only 24 evaluators discussed involving participants in the evaluation when asked how age affects their evaluation, 20 evaluators discussed how they would not survey individuals under a certain age, 10 discussed how they would not interview individuals under a certain age, and seven discussed how they would only observe or use other indirect methods with individuals under a certain age. Given the importance of considering the developmental context for conducting culturally competent evaluation (AEA, 2011; Choinard & Cousins, 2009), this suggests that evaluators may not be considering age as an important cultural context to consider when designing and implementing evaluations.

Beyond participatory evaluation practices, there were few differences in the practices of evaluators between age conditions. There were differences in the administration techniques used, with evaluators in the child condition more likely to have program staff administer surveys, use oral survey administration, and use interview specialists and less likely to use online survey administration or use the evaluation team to conduct interviews. Evaluators in the child condition were also more likely to use shorter interviews, fewer participants in focus groups, and comprise focus groups of participants of similar ages, indicating evaluators are thinking critically about developmental appropriateness in the context of interviews and focus groups.

Some of the lack of differences between age conditions do not necessarily mean lack of developmentally appropriate practice, however. For instance, evaluators did not differ in evaluation design between conditions and nearly all evaluators used a longitudinal approach to the three-year evaluation. While evaluations involving young populations should perhaps use designs that better account for maturation effects (i.e., quasi-experiments and experiments), evaluations are not restricted to studying outcomes and should also explore fidelity of implementation (McCall & Green, 2004). Evaluators also did not differ in the number of

response options they would use on a survey, with nearly all evaluators choosing between 2-7 response options; however, studies have shown that children respond better to surveys with 2-4 response options instead of 5-7. Furthermore, evaluators did not differ in the structure of the interview or focus group protocols, with the clear majority choosing semi-structured protocols, and did not differ in the types of analyses they would conduct, indicating evaluators may have preferred data analytic techniques regardless of the evaluation context.

One way evaluators can promote developmentally appropriate practice in the context of youth programs is by including developmental expertise on the evaluation team (Bialosiewicz, 2013). However, evaluators in the child or adolescent conditions were not more likely to include developmental expertise on the evaluation team compared to evaluators in the adult condition. This may have been due to levels of personal developmental expertise that evaluators already have; thus, membership in youth-related TIGs (i.e., PreK-12 Education TIG and Youth-Focused Evaluation TIG) as well as self-reported knowledge and experience working with youth populations was examined as a possible moderator of developmentally appropriate practice. However, there were no differences in conditions between evaluators with varying levels of developmental expertise and in fact evaluators with more developmental expertise were less likely to be participatory. Even other evaluator characteristics (i.e., education level, evaluator role, years as an evaluator) did not differ between age conditions.

These results leave us with two questions: (a) which evaluators are more likely to have developmentally appropriate practice and (b) what are the barriers to developmentally appropriate practice in the context of youth program evaluation? There is evidence that even evaluators experienced in working with youth may need help in conducting developmentally appropriate evaluations (Bialosiewicz, 2015). In interviews with youth evaluators, many demonstrated strong developmentally appropriate practice (e.g., nearly all the evaluators

regularly collected data from children and youth) yet felt there were barriers to high quality youth program evaluation practice. These barriers included cost and time needed and misconceptions from clients about good evaluation practice, while a few evaluators mentioned that they felt inhibited by their lack of developmental expertise or methodological skills. These may be issues faced by many youth program evaluators, suggesting a need or more training in developmentally appropriate practice or better resources (i.e., time and money) for conducting youth program evaluations.

Limitations and Future Directions

Some limitations of this study should be discussed. First, despite simulation studies mimicking a real-world experience, participants can make “riskless” choices that do not lead to real world consequences (Ward, 1954) and, with the high number of closed-ended questions to reduce survey fatigue, there is concern for social desirability. However, research has suggested simulation studies approximate other methodological approaches well (Cannon & Burns, 1999; O’Neil, Allred, & Dennis, 1997; Smith, Winer, & George, 1983) and this study analyzed the content of descriptions, evaluations, and RFPs of tutoring programs, enabling this study to better simulate a real-world program and evaluation experience. Second, many of the participating evaluators may not have had experience working with youth programs, or with tutoring programs. This limitation was hopefully lessened through describing the specific outcomes and implementation the program was seeking; however, some evaluators may feel uncomfortable evaluating a type of program they are unfamiliar with. Evaluators with a wide range of experience were sought specifically to determine how experience with the type of program (i.e., experience working with child and adolescent serving programs) affects the evaluation approach, design, and methods used. Third, with the low response rate (27%), the findings of this study may not be generalizable to the entire population of AEA members, or to evaluators in general,

due to self-selection bias.

Future studies should continue this work on developmentally appropriate evaluation practice. Replication studies and other similar studies should be conducted to come to a greater consensus on the prevalence of developmentally appropriate evaluation practice. This study only briefly touched upon participatory evaluations with children and adolescents, and more work in this area should be conducted. This study also purposefully avoided discussion of ethics (i.e., IRB and FERPA guidelines, protecting children and adolescents as vulnerable populations) to narrow the scope of the study, but this topic appeared in the open-ended responses to questions about how age affects the evaluation process in general. Lastly, given how many of the issues in developmentally appropriate practice involved data collection techniques, particularly in survey methodology, more research should be conducted on specific methodological techniques with children and adolescents, such as how to write survey or interview questions, how to design and analyze surveys, and how to conduct appropriate interviews or focus groups.

References

- American Evaluation Association. (2004). American Evaluation Association guiding principles for evaluators. Fairhaven, MA: American Evaluation Association. Retrieved from <http://www.eval.org/Publications/GuidingPrinciples.asp>.
- American Evaluation Association. (2011). American Evaluation Association public statement on cultural competence in evaluation. Fairhaven, MA: American Evaluation Association. Retrieved from <http://www.eval.org/ccstatement.asp>.
- Azzam, T. (2007). *Evaluator contextual responsiveness*. Doctoral dissertation. Retrieved from Proquest: 3272251.
- Azzam, T. (2011). Evaluator characteristics and methodological choice. *American Journal of Evaluation* 32 (3), 376-391.
- Bell, A. (2007). Designing and testing questionnaires for children. *Journal of Research in Nursing*, 12(5), 461–469.
- Berry, T., Menkes, S. & Bono, K. (2009). Improving evaluation practice through the developmental context. Paper presented at the annual meeting of the American Evaluation Association, Orlando FL.
- Bialosiewicz, S. (2013). *Current trends in youth-program evaluation*. Unpublished manuscript, Department of Behavioral and Organizational Sciences, Claremont Graduate University, Claremont, CA.
- Better Evaluation (2016). *Specify the key evaluation questions*. Retrieved from http://betterevaluation.org/en/plan/engage_frame/decide_evaluation_questions
- Borgers, N., & Hox, J. (2001). Item nonresponse in questionnaire research with children. *Journal of Official Statistics*, 17(2), 321–335.
- Borgers, N., de Leeuw, E., & Hox, J. (2000). Children as respondents in survey research:

- Cognitive development and response quality 1. *Bulletin de Methodologie Sociologique*, 66(1), 60–75.
- Borgers, N., Hox, J., & Sikkels, D. (2003). Response quality in survey research with children and adolescents: The effect of labeled response options and vague quantifiers. *International Journal of Public Opinion Research*, 15(1), 83–94.
- Bronfenbrenner, U., & Morris, P. A. (2006). The bioecological model of human development. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology, Vol. 1: Theoretical models of human development* (6th ed., pp. 793-828). New York: John Wiley.
- Cannon, H. M., & Burns, A. C. (1999). A framework for assessing the competencies reflected in simulation performance. *Developments in Business Simulation & Experiential Exercises*, 26, 40-44.
- CASEL – Collaborative for Academic, Social, and Emotional Learning (2012). *Effective social and emotional learning programs*. Chicago, IL: Author.
- Chen, H. T. (2005). *Practical program evaluation: Assess and improve program planning, implementation, and effectiveness*. Thousand Oaks, CA: Sage.
- Chouinard, J. A., & Cousins, J. B. (2009). A review and synthesis of current research on cross-cultural evaluation. *American Journal of Evaluation*, 30(4), 457-494.
- Christensen, P., & James, A. (2008). *Research with children: Perspectives and practices*. Routledge.
- Christensen, P., & Prout, A. (2002). Working with ethical symmetry in social research with children. *Childhood*, 9(4), 477–497.
- Christie, C.A. (2003). What guides evaluation? A study of how evaluation practice maps onto evaluation theory. In C.A. Christie (Ed.), *The practice-theory relationship*. (New Directions for Program Evaluation, No. 97) (pp. 7–36). San Francisco, CA: Jossey-Bass.

- Christie, C.A., & Fleischer, D. (2010). Insight into evaluation practice: An analysis of evaluation studies published in North American evaluation focused journals. *American Journal of Evaluation* 31(3), 326-346.
- Clark, L. (2009). Focus group research with children and youth. *Journal for Specialists in Pediatric Nursing*, 14(2), 152.
- Coad, J., & Evans, R. (2007). Reflections on practical approaches to involving children and young people in the data analysis process: Involving children in the data analysis process. *Children & Society*, 22(1), 41–52. <http://doi.org/10.1111/j.1099-0860.2006.00062.x>
- Coryn, C.L., Ozeki, S., Wilson, L.N., Greenman, G.D., Schroter, D.C., Hobson, K.A., ... Vo, A.T. (2015). Does research on evaluation matter? Findings from a survey of American Evaluation Association members and prominent evaluation theorists and scholars. *American Journal of Evaluation*, 37(1), 2016.
- Crano, W. D., & Brewer, M. B. (2002). *Principles and Methods of Social Science Research* (2nd Ed.). New York: Psychology Press.
- Daigneault, P., & Jacob, S. (2009). Toward accurate measurement of participation: Rethinking the conceptualization and operationalization of participatory evaluation. *American Journal of Evaluation*, 30, 330-348.
- Davis, J. M. (1998). Understanding the meanings of children: A reflexive process. *Children & Society*, 12(5), 325–335.
- de Leeuw, E. (2011). *Improving data quality when surveying children and adolescents: Cognitive and social development and its role in questionnaire construction and pretesting*. Finland: Report prepared for the Annual Meeting of the Academy of Finland: Research Programs Public Health Challenges and Health and Welfare of Children and Young People.

- de Leeuw, E., Borgers, N., & Smits, A. (2004). Pretesting questionnaires for children and adolescents. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 409–429). Hoboken, NJ: John Wiley & Sons, Inc.
- Donaldson, S. I., Christie, C. A., & Mark, M. M. (Eds.) (2009). *What counts as credible evidence in applied research and evaluation practice?*. Los Angeles, CA: Sage.
- Eder, D., & Fingerson, L. (2003). Interviewing children and adolescents. In J. A. Holstein & J. F. Gubrium (Eds.), *Inside interviewing: New lenses, new concerns*.
- Eide, B. & Winger, N. (2005). From the children's point of view: Methodological and ethical challenges. In A. Clark, A. Kjørholt, & P. Moss (Eds.), *Beyond Listening: Children's perspectives on Early Childhood Services*. Bristol: Policy Press.
- Einarsdóttir, J. (2007). Research with children: methodological and ethical challenges. *European Early Childhood Education Research Journal*, 15(2), 197–211.
<http://doi.org/10.1080/13502930701321477>
- Elder, G. H. (1998). The life course as developmental theory. *Child Development*, 69(1), 1-12.
- Fargas-Malet, M., McSherry, D., Larkin, E., & Robinson, C. (2010). Research with children: Methodological issues and innovative techniques. *Journal of Early Childhood Research*, 8(2), 175–192.
- Fine, G. A., & Sandstrom, K. L. (1988). *Knowing children: Participant observation with minors*. London: SAGE Publications Ltd.
- Fitzpatrick, J.L, Sanders, J.R., & Worthen, B.R. (2004). Program evaluation: Alternative approaches and practical guidelines (3rd edition). Upper Saddle River, NJ: Pearson Education Inc.
- Gallagher, M. (2009). Data collection and methods. In E. K. M. Tisdall, J. M. Davis, & M.

- Gallagher, *Researching with children and young people: Research design, methods, and analysis*. London: SAGE Publications Ltd.
- Gibson, F. (2007). Conducting focus groups with children and young people: Strategies for success. *Journal of Research in Nursing* 12(5), 473-483.
- Gong, J. & Wright, D. (2007). The context of power: Young people as evaluators. *American Journal of Evaluation*, 28(3), 327-333.
- Graue, M. E., & Walsh, D. J. (1998). *Studying children in context : Theories, methods, and ethics*. Thousand Oaks, Calif: SAGE Publications, Inc.
- Hart, R.A. (1992). Children's participation: From tokenism to citizenship. UNICEF Innocenti Essays, No. 4, Florence, Italy: International Child Development Centre of UNICEF.
- Heary, C.M., & Hennessy, E. (2002). The use of focus group interviews in pediatric health care research. *Journal of Pediatric Psychology*, 27(1), 47-57
- Hill, M. (1997). Participatory research with children. *Child and Family Social Work*, 2, 171-183.
- Jacobs, F. (2003). Child and family program evaluation: Learning to enjoy complexity. *Applied Developmental Science*, 7(2), 62-75.
- Kirby, P. (1999) *Involving Young Researchers*. York: York Publishing Services.
- Larsson, A. S., & Lamb, M. E. (2009). Making the most of information-gathering interviews with children. *Infant and Child Development*, 18(1), 1-16. <http://doi.org/10.1002/icd.573>
- Lerner, R. M. (2006). Developmental science, developmental systems, and contemporary theories of human development. In R. M. Lerner, W. Damon (Eds.), *Handbook of child psychology (6th ed.): Vol 1, Theoretical models of human development* (pp. 1-17). Hoboken, NJ: John Wiley & Sons Inc.
- Lewis, A. (1992). Group child interviews as a research tool. *British Educational Research Journal*, 18(4), 413-421.

- Mac Naughton, G., Rolfe, S. A., & Siraj-Blatchford, I. (Eds.). (2001). *Doing early childhood research: International perspectives on theory and practice*. Buckingham: Open University Press.
- Mauthner, M. (1997). Methodological aspects of collecting data from children: Lessons from three research projects. *Children & Society*, 11(1), 16–28.
- McCall, R. B., & Green, B. L. (2004). Beyond the methodological gold standards of behavioral research: Considerations for practice and policy. *Social Policy Report*, 18(2).
- Morrow, V., & Richards, M. (1996). The ethics of social research with children: An overview. *Children & Society*, 10, 90–105.
- O'Neil, H. F., Allred, K. G., & Dennis, R. (1997). Validation Studies of a Computer Simulation Measure of Negotiation Skills. In H. F. O'Neil (Ed.), *Workforce Competencies and Assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.
- O'Donnell, C.L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K -12 curriculum intervention research. *Review of Educational Research*, 78, 33-84.
- Ogan, C., Karakuş, T., & Kurşun, E. (2013). Methodological issues in a survey of children's online risk-taking and other behaviours in Europe. *Journal of Children and Media*, 7(1), 133–150. <http://doi.org/10.1080/17482798.2012.739812>
- Parkinson, D. D. (2001). Securing trustworthy data from an interview situation with young children: Six integrated interview strategies. *Child Study Journal*, 31(3), 137–156.
- Petrosino, A. (2003). Estimates of randomized controlled trials across six areas of childhood intervention: A bibliometric analysis. *The Annals of the American Academy of Political and Social Science*, 589(1), 190-202.
- Powers, J.L., & Tiffany, J.S. (2006). Engaging youth in participatory research and evaluation.

- Journal of Public Health Management and Practice*, November (Supplemental), 79-87.
- Punch, S. (2002). Research with children: The same or different from research with adults? *Childhood*, 9(3), 321-341.
- Rapport, M. D., Bolden, J., & Chung, K. (2011). Issues in doing research with children. In J. C. Thomas & M. Hersen (Eds.), *Understanding research in clinical and counseling psychology*. Taylor & Francis.
- Rossi, P. H., Freeman, H. E., & Lipsey, M. W. (1999). *Evaluation: A systematic approach*. Thousand Oaks, Calif: Sage Publications.
- Sabo Flores, K. (2008). *Youth participatory evaluation: Strategies for engaging young people*. San Francisco: John Wiley & Sons.
- Scott-Little, C., Hamann, M. S., & Jurs, S. G. (2002). Evaluations of after-school programs: A meta-evaluation of methodologies and narrative synthesis of findings. *American Journal of Evaluation*, 23(4), 387-419.
- Scott, J. (2008). Children as respondents: The challenge for quantitative methods. In P. Christensen & A. James (Eds.), *Research with children: Perspectives and practices*. New York, NY: Routledge.
- Shonkoff, J. P., & Phillips, D. A. (Eds.). (2000). Promoting healthy development through intervention. In *From neurons to neighborhoods: The science of early childhood development*, (pp. 337-380). Washington D.C.: Academy Press.
- Smith, H. W., Winer, J. L., & George, C. E. (1983). The relative efficacy of simulation experiments. *Journal of Vocational Behavior*, 22(1), 96-104.
- Tisdall, E. K. M., Davis, J. M., & Gallagher, M. (2009). *Researching with children and young people: Research design, methods, and analysis*. London: SAGE Publications Ltd.
- Tudge, J. R. H., Mokrova, I. Hatfield, B. E., & Karnik, R. B. (2009). Uses and misuses of

- Bronfenbrenner's bioecological theory of human development. *Journal of Family Theory & Review*, 1, 198-210.
- Vaughn, S., Schumm, J.S., & Sinagub, J. (1996). *Focus group interviews in education and psychology*. Thousand Oaks, CA: Sage.
- Weiss, C. H. (1998). *Evaluation*, (2nd ed.). Upper Saddle River, NJ: Prentice Hall.
- Weithorn, L. A., & Scherer, D. G. (1994). Childrens involvement in research participation decisions: Psychological considerations. In M. A. Grodin & L. H. Glantz, *Children as research subjects: Science ethics, and law*. New York, NY: Oxford University Press.
- Wholey, J. S., Hatry, H. P., & Newcomer, K. E. (2004). *Handbook of practical program evaluation*. San Francisco, CA: Jossey-Bass.
- Woodhead, M., & Faulkner, D. (2008). Subjects, objects, or participants? Dilemmas of psychological research with children. In P. Christensen & A. James (Eds.), *Research with children: Perspectives and practices*. New York, NY: Routledge.