

High/Scope Youth PQA Technical Report



Inter-rater Reliability on the Youth Program Quality Assessment

Juliane Blazevski and Charles Smith
High/Scope Educational Research Foundation



Introduction

The Youth Program Quality Assessment (PQA) is an observational assessment tool developed by the High/Scope Educational Research Foundation. It is used to evaluate the quality of youth programs and identify staff development needs. The Youth PQA assesses a broad range of best practices that are applicable in many settings, regardless of a program’s content. The Youth PQA is made up of two forms: Form A, which focuses on youth experiences during a program offering, and Form B, which focuses on the organization’s infrastructure. A complete list of scales, subscales, and items are presented in the appendix. The Form A of the Youth PQA has already been the subject of a validation study (Smith & Hohmann, 2005); however, this technical report will focus specifically on inter-rater reliability findings assembled since the publication of the original study.

Data Collection Procedures

Completion of Form A requires a 2-hour observation of a specific program offering (e.g., book club, basketball, arts and crafts) at a youth program site. A trained external rater (data collector) gathers anecdotal evidence and records it in the space provided on the form or on a separate sheet of paper. This evidence might include descriptions of interactions between staff and youth, quotations of what youth and staff say, actions of staff and youth, materials lists, and sequences of events or activities. To provide an overall rating of offering-level quality in an organization, multiple Form As are typically completed (one for each program offering or a random selection of program offerings) and then these individual ratings are combined (averaged).

After the observation period is complete, the rater uses the anecdotal evidence to score the items (behavioral indicators) associated with various facets of best practice. Items are scored on a 3-point scale scored 1, 3, 5, or NR (not rated), with a score of 1 corresponding to a low score and a 5 corresponding to a high score for the best practice described in the item. Scores of 2 and 4 are not allowed.¹ An example subscale, “Staff support youth in building new skills,” is presented to the right.

Staff support youth in building new skills (subscale)			Supporting Evidence/Anecdotes
Items			
1 Youth are not encouraged to try out new skills or attempt higher levels of performance.	3 Some youth are encouraged to try out new skills or attempt higher levels of performance but others are not.	5 All youth are encouraged to try out new skills or attempt higher levels of performance.	<input type="checkbox"/>
1 Some youth who try out new skills with imperfect results, errors, or failure are informed of their errors (e.g., “That’s wrong”) and/or are corrected, criticized, made fun of, or punished by staff <i>without</i> explanation.	3 Some youth who try out new skills receive support from staff who problem-solve with youth despite imperfect results, errors, or failure, and/or some youth are corrected <i>with</i> an explanation.	5 All youth who try out new skills receive support from staff despite imperfect results, errors, or failure; staff allow youth to learn from and correct their own mistakes and encourage youth to keep trying to improve their skills.	<input type="checkbox"/>

Training Methodology

High/Scope provides a 2-day training sequence to support use of the Youth PQA–Form A. Day 1 is focused on getting participants familiar with standard observational protocol and collection of objective anecdotal data, and provides opportunities to discuss the intent of every item on Form A. Participants

¹ The items were initially scored with a 1, 2, or 3, but because of concerns that a 1-point difference in a scale score may be interpreted by practitioners as not a meaningful difference, the range of the rating scale was expanded to better capture the degree of difference between a low and a mid score and between a mid and high score for each item.

fit premade anecdotes — short fictional scenarios related to specific areas of quality — to appropriate items and score them. At the end of Day 1, levels of inter-rater agreement are calculated for the entire training group based on these scores. Day 1 training emphasizes areas in which raters disagree to build shared understanding of the items. Day 2 deepens participants’ understanding of Form A and the process of anecdotal evidence gathering and lets participants practice using the Youth PQA on a series of short video clips. Participants are required to achieve 80% perfect agreement with the expert item-level scores for these video segments to be certified as reliable Youth PQA data collectors. Individuals who have already completed the 2-day training but wish to refresh their skills are encouraged to complete a 1-day training, which involves a review of the observational protocol and a reliability test based on the scoring of a 30-minute video segment of a youth program offering. As with the 2-day training, participants are required to achieve 80% perfect agreement with the “expert” item-level scores for these video segments to be certified as reliable.

Inter-rater Reliability

Background

The Youth PQA Validation Study (Smith & Hohmann, 2005) reported prior work examining inter-rater reliability on an earlier version of the Youth PQA. In this study, the researchers examined inter-rater reliabilities for the four main scales on Form A: *safe environment*, *supportive environment*, *interaction*, and *engagement*. They examined raters’ agreement on observations by using a statistic known as the *intraclass correlation (ICC)*, which examines the degree to which differences among all ratings have to do with the difference between raters or the differences among the programs themselves. A high ICC means that there is more variation across offerings than within raters and indicates a high degree of inter-rater agreement. Youth PQA–Form A scales yielded acceptable² inter-rater reliability with the exception of the safe environment scale. ICCs for paired-raters on the four scales were as follows: *safe environment* = .48, *supportive environment* = .69, *interaction* = .83, and *engagement* = .72.

Unfortunately, intraclass correlations are difficult to understand for nontechnical readers and have not been the inter-rater statistic of choice in recent work on the quality assessments (see discussion in Yohalem & Wilson-Ahlstrom, 2007). To support both ease of understanding and cross-instrument performance comparisons, this paper presents new inter-rater findings for the Youth PQA in two forms: percent perfect agreement at the lowest level (item) of measurement and Kappa. Kappa statistics are often described as being a “chance corrected” measure of inter-rater reliability in that they are a ratio of agreements to disagreements in relation to expected frequencies, although there are some well-known problems with necessary assumptions for this interpretation.³ Recent analyses presented in the following sections of this report indicate that the current version of the assessment tool, paired with improved training techniques, produces acceptable levels of inter-rater reliability.

Methods

In the past 2 years, High/Scope researchers have captured four paired-rater data sets for a total of 32 rater pairs. Twenty-two of the pairs observed live program offerings, and 10 observed a 30-minute video segment of a program offering. All raters used the current version of the Youth PQA–Form A. When investigating inter-rater reliability, we examined item-level scores (subscales on the Youth PQA

² ICC > .70 is generally considered an indication of acceptable levels of inter-rater agreement, although there is much debate about the use of arbitrary cutoffs for this index (Harvey & Hollander, 2004; James, Demaree, & Wolf, 1984).

³ See Uebersax (1987) for a discussion of the appropriateness of the Kappa coefficient for quantifying levels of agreement, including a discussion of the violation of the assumption that raters are statistically independent and the role of chance on the decisions of raters.

have 2–6 items that are rated by the data collectors and then summed to produce subscale scores). Computing inter-rater reliability at the lowest level of the instrument is critical since these item scores represent the actual behaviors observed and scored. Raters were only said to have agreed if the scores they gave to a particular item were identical. In other words, we used “perfect agreement” as our standard. Because we used the most stringent method for computing reliability, as opposed to methods that focus on agreement for scale-level scores in which slight differences in actual ratings are essentially averaged out or methods that allow a 1-point difference in scores to still count as agreement, caution should be used when interpreting the scores, particularly in comparison to other observational quality instruments. That is, one should not assume that the Youth PQA is less reliable than another instrument just because the reported Kappa is lower without examining the level of analysis and determining the criteria used for agreement.

Results

Our most recent inter-rater reliability findings are presented in Table 1. The table presents item-level perfect agreement and the corresponding Kappa coefficient. Sections to the right present the average percent agreement and Kappa for all items within a subscale, scale, and for the entire Youth PQA–Form A. Landis and Koch (1977) suggest that Kappa statistics between .40 and .59 indicate moderate agreement and Kappa statistics between .60 and .79 indicate substantial agreement. We found that across the 32 rater pairs there was 78% perfect agreement at the item level, yielding an overall Kappa coefficient of .67 for Form A, indicating substantial overall agreement for items on this instrument. Looking more closely at the distribution of Kappa scores across the items, we found that 68% of the items produced Kappas within the substantial agreement range, 27% of the items produced Kappas within the moderate agreement range, and only 5% of the items produced Kappas below this range (poor agreement). Kappa coefficients for the four subscales⁴ range from 0.54 to 0.73. It is important to note, however, that the items with lower reliability were often not rated/observed in all 32 paired observations (smaller sample size). For example, subscale K pertains to conflict resolution. If no conflicts occurred, then items on this subscale were not scored and therefore could not be used in these analyses.

Discussion

The results reported above suggest that the current version of the Youth PQA–Form A, paired with improved training techniques, produces acceptable levels of inter-rater reliability. In other words, when two or more raters observe the same youth program offering at the same time, they produce identical item scores in most cases (78% perfect agreement). The results also identified areas for instrument revision, particularly on the items comprising the interaction scale and more specifically on subscale O (Youth have opportunities to partner with adults). However, even without further revision, one can assume that differences in program quality assessed by this tool reflect real differences between programs and not just incidental differences between raters.

⁴ Kappa coefficients reflect item-level agreement for items comprising a particular subscale, not rater agreement on a computed subscale score.

Table 1. Inter-rater reliability for Youth PQA-Form A: Percent perfect agreement and Kappa coefficient (N = 32 pairs)

Item	Grand Total	Total Possible	% Perfect Agreement (item-level)	Kappa	Subscale	% Perfect Agreement (item-level)	Kappa	Scale	% Perfect Agreement (item-level)	Kappa
I.A1	26	32	81.3%	0.72	I.A	78.1%	0.67	SAFETY	82.2%	0.73
I.A2	24	32	75.0%	0.63	I.B	83.6%	0.75	SUPPORT	79.8%	0.70
I.B1	22	32	68.8%	0.53	I.C	73.0%	0.59	INTERACTION	69.1%	0.54
I.B2	26	32	81.3%	0.72	I.D	86.4%	0.80	ENGAGEMENT	77.1%	0.66
I.B3	29	32	90.6%	0.86	I.E	77.7%	0.67			
I.B4	30	32	93.8%	0.91	II.F	88.9%	0.83			
I.C1	31	32	96.9%	0.95	II.G	81.0%	0.72	FORM A	78.0%	0.67
I.C2	18	22	81.8%	0.73	II.H	84.4%	0.77			
I.C3	15	17	88.2%	0.82	II.I	75.0%	0.63			
I.C4	2	3	66.7%	0.50	II.J	75.0%	0.63			
I.C5	15	22	68.2%	0.52	II.K	69.1%	0.54			
I.C6	10	15	66.7%	0.50	III.L	66.4%	0.50			
I.D1	27	32	84.4%	0.77	III.M	80.2%	0.70			
I.D2	28	32	87.5%	0.81	III.N	68.4%	0.53			
I.D3	26	29	89.7%	0.84	III.O	56.6%	0.35			
I.D4	27	32	84.4%	0.77	IV.P	81.8%	0.73			
I.E1	25	32	78.1%	0.67	IV.Q	78.1%	0.67			
I.E2	25	31	80.6%	0.71	IV.R	75.0%	0.63			
I.E3	23	31	74.2%	0.61						
II.F1	21	26	80.8%	0.71						
II.F2	31	32	96.9%	0.95						
II.F3	28	32	87.5%	0.81						
II.G1	29	32	90.6%	0.86						
II.G2	25	31	80.6%	0.71						
II.G3	27	31	87.1%	0.81						
II.G4	25	32	78.1%	0.67						
II.G5	22	32	68.8%	0.53						
II.H1	27	32	84.4%	0.77						
II.H2	28	32	87.5%	0.81						
II.H3	26	32	81.3%	0.72						
II.H4	26	32	81.3%	0.72						
II.I1	25	32	78.1%	0.67						
II.I2	23	32	71.9%	0.58						
II.J1	27	32	84.4%	0.77						
II.J2	22	32	68.8%	0.53						
II.J3	23	32	71.9%	0.58						
II.K1	14	17	82.4%	0.74						
II.K2	9	17	52.9%	0.29						
II.K3	11	17	64.7%	0.47						
II.K4	13	17	76.5%	0.65						
III.L1	23	32	71.9%	0.58						
III.L2	21	32	65.6%	0.48						
III.L3	23	32	71.9%	0.58						
III.L4	18	32	56.3%	0.34						
III.M1	27	32	84.4%	0.77						
III.M2	27	32	84.4%	0.77						
III.M3	23	32	71.9%	0.58						
III.N1	20	32	62.5%	0.44						
III.N2	13	17	76.5%	0.65						
III.N3	19	27	70.4%	0.56						
III.O1	18	32	56.3%	0.34						
III.O2	12	21	57.1%	0.36						
IV.P1	18	22	81.8%	0.73						
IV.P2	18	22	81.8%	0.73						
IV.Q1	25	32	78.1%	0.67						
IV.Q2	25	32	78.1%	0.67						
IV.R1	27	32	84.4%	0.77						
IV.R2	25	32	78.1%	0.67						
IV.R3	23	32	71.9%	0.58						
IV.R4	21	32	65.6%	0.48						
TOTALS:	1337	1715								

Kappa = (% perfect agreement - chance agreement) / (1- chance agreement)

"chance agreement" = .333 assuming equal cell counts across 9 cells (3X3 crosstabs of Rater 1 item score with Rater 2 item score)

References

- Harvey, R., & Hollander, E. (2004, April). *Benchmarking r_{ng} interrater agreement indices: Let's drop the .70 rule-of-thumb*. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL.
- James, L., Demaree, R., & Wolf, G. (1984). Estimating within-group inter-rater reliability with and without response bias. *Journal of Applied Psychology, 69*(1), 85–98.
- Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.
- Smith, C., & Hohmann, C. (2005). *Youth Program Quality Assessment validation study: Findings for instrument validation*. High/Scope Press: Ypsilanti, MI.
- Uebersax, J. (1987). Diversity of decision-making models and the measurement of inter-rater agreement. *Psychological Bulletin, 101*, 140–146.
- Yohalem, N., & Wilson-Ahlstrom, A. (2007). *Measuring youth program quality: A guide to assessment tools*. Washington, DC: Forum for Youth Investment.

Appendix

Youth PQA – Form A (Version 5.0)	
I. SAFE ENVIRONMENT	II. SUPPORTIVE ENVIRONMENT
<p>A. Psychological and emotional safety.</p> <ol style="list-style-type: none"> 1. Emotional climate 2. Mutual respect (religion, ethnicity, etc.) <p>B. Physically safe environment.</p> <ol style="list-style-type: none"> 1. Health and safety 2. Sanitation 3. Ventilation and lighting 4. Temperature <p>C. Emergency procedures and supplies.</p> <ol style="list-style-type: none"> 1. Emergency procedures 2. Fire extinguisher 3. First aid kit 4. Other safety equipment 5. Supervised entrances 6. Supervised access to outdoor space <p>D. Program space and furniture.</p> <ol style="list-style-type: none"> 1. Sufficient space 2. Suitable space 3. Furniture 4. Physical environment can be modified <p>E. Healthy food and drinks are provided.</p> <ol style="list-style-type: none"> 1. Drinking water 2. Available food and drinks 3. Healthy food and drinks 	<p>F. Staff provide a welcoming atmosphere.</p> <ol style="list-style-type: none"> 1. Staff greet youth 2. Staff tone of voice and language 3. Staff smile, friendly gestures, eye contact <p>G. Session flow.</p> <ol style="list-style-type: none"> 1. Start and end on time 2. Materials and supplies ready 3. Enough materials and supplies for all youth 4. Staff explain activities clearly 5. Appropriate time for activities <p>H. Activities support active engagement.</p> <ol style="list-style-type: none"> 1. Youth engage with materials or ideas 2. Tangible products or performances 3. Youth talk about what they are doing 4. Balance concrete and abstract <p>I. Staff support youth in building new skills.</p> <ol style="list-style-type: none"> 1. Youth encouraged to try new skills 2. Mistakes allowed <p>J. Staff support youth with encouragement.</p> <ol style="list-style-type: none"> 1. Staff actively involved with youth 2. Staff use specific, nonevaluative language 3. Open-ended questions <p>K. Manage feelings and resolve conflicts.</p> <ol style="list-style-type: none"> 1. Acknowledge feelings 2. Help youth respond appropriately 3. Ask youth what happened 4. Ask for solutions
III. INTERACTION	IV. ENGAGEMENT
<p>L. Sense of belonging.</p> <ol style="list-style-type: none"> 1. Get to know each other 2. Inclusive relationships 3. Youth identify with program offering 4. Publicly acknowledge achievements <p>M. Small groups.</p> <ol style="list-style-type: none"> 1. Groupings 2. Ways to form small groups 3. Groups have purpose and cooperation <p>N. Youth facilitators and mentors.</p> <ol style="list-style-type: none"> 1. Group-process skills 2. Opportunities to mentor 3. Opportunities to lead a group <p>O. Partner with adults.</p> <ol style="list-style-type: none"> 1. Staff share control with youth 2. Staff provide an explanation 	<p>P. Set goals and make plans.</p> <ol style="list-style-type: none"> 1. Plans for projects and activities 2. Planning strategies <p>Q. Make choices based on their interests.</p> <ol style="list-style-type: none"> 1. Content choices 2. Process choices <p>R. Youth have opportunities to reflect.</p> <ol style="list-style-type: none"> 1. Youth reflect on what they are doing 2. Youth reflect in multiple ways 3. Youth make presentations 4. Youth give feedback on the activities