Appendix

Appendix A1 Study characteristics: Schultz, Barr, & Selman, 2001 (quasi-experimental design)

Characteristic	Description
Study citation	Schultz, H. L., Barr, D. J., & Selman, R. L. (2001). The value of a developmental approach to evaluating character development programmes: An ongoing study of Facing History and Ourselves. *Journal of Moral Education,* 30, 3–25.
Participants	The study included 346 eighth-grade public school students from 14 FHAO and 8 comparison classrooms in social studies and language arts. The sample was 62% white, 6% black, 3.5% Hispanic, and 23% mixed/other students, with 5.5% of the students not reporting their ethnicity.
Setting	The participating classrooms were in northeastern U.S. towns with varied socioeconomic characteristics: a suburban town with middle class and wealthy families, an urban suburb with a mix of wealthy, middle class, and working class families, and two small cities with predominantly poor and working class families.
Intervention	Students in 14 eighth-grade classrooms taught by four teachers with experience implementing *Facing History and Ourselves* used the curriculum over a 10-week period. Information on the specific FHAO curriculum they used was not provided in the study report, and the authors note that the curriculum generally varies in length and content. The core components include readings from the *Facing History and Ourselves* resource book, guest speakers, films, and student writings around such themes as morality, justice, and caring for others.
Comparison	Students in eight eighth-grade classrooms taught by five teachers in public schools in the same communities as the FHAO teachers but, with one exception, not in the same schools as the FHAO teachers.
Primary outcomes and measurement	The primary outcomes included self-reported fighting, relationship maturity, ethnic identity, civic attitudes and participation, racism, and moral reasoning. Self-reported fighting was measured with a questionnaire, but no other details were provided. Relationship maturity was measured with The Group for the Study of Interpersonal Development relationship questionnaire. The measure of ethnic identity was the Multi-group Ethnic Identity Measure. Civic attitudes and participations were assessed with scales adapted from the National Learning Through Service Survey. The Modern Racism Scale measured racism, and the Defining Issues Test, moral reasoning. (See Appendices A2.1 and A2.2.)
Teacher training	Training was one of the selection criteria for intervention group teachers. Each teacher for the FHAO classes had attended the FHAO Institute and had taught the FHAO curriculum for at least three years before the study.

Appendix A2.1 Outcome measures in the behavior domain

Outcome measure	Description
Fighting	Students' self-reported fighting behavior (as cited in Schultz, Barr, & Selman, 2001).

Appendix A2.2 Outcome measures in the knowledge, attitudes, and values domain

Outcome measure	Description
GSID Relationship Questionnaire: Relationship maturity (best response score) scale	GSID Relationship Questionnaire (as cited in Schultz, Barr, & Selman, 2001): Relationship Maturity, scored by "best response." This questionnaire, developed by the Group for the Study of Interpersonal Development, has 24 multiple-choice items. Five scales from these 24 items are averaged for the total score. The best response score is based on the student's choice of the best response of four possible responses to each question.
GSID Relationship Questionnaire: Relationship maturity (response rating score) scale	GSID Relationship Questionnaire (as cited in Schultz, Barr, & Selman, 2001): Relationship Maturity, scored by "response rating." This instrument, described above, uses the same items but different response options. The response rating score is based on the student's assignment of "poor," "average," "good," and "excellent" to each of four responses to each question.
McConahay Modern Racism scale	McConahay Modern Racism scale (as cited in Schultz, Barr, & Selman, 2001).
Phinney Multigroup Ethnic Identity Measure	Phinney Multigroup Ethnic Identity Measure (MEIM, as cited in Schultz, Barr, & Selman, 2001). The 14 items on this measure make up three scales, which are averaged into the total score.
Civic attitudes and participation	Scales adapted from the National Learning Through Service Survey developed by the Search Institute (as cited in Schultz, Barr, & Selman, 2001). The six subscales used in this study were feelings about intergroup differences, beliefs about civic responsibility, importance of civic activism, involvement with social issues, anticipated future activism, and sense of agency. These scores were averaged together to construct an overall civic attitudes and participation scale.
Defining Issues Test: Moral reasoning (P score)	Defining Issues Test (DIT, as cited in Schultz, Barr, & Selman, 2001), P Score (% of principled moral reasoning in responses). In this test, students rate 12 statements, which are based on four dilemmas. The P score indicates the importance that students place on principled moral considerations in making a moral decision.
Defining Issues Test: Moral reasoning (D score)	Defining Issues Test (DIT, as cited in Schultz, Barr, & Selman, 2001), D score (composite moral reasoning score). The test is the same as that for the P score, but the D score is more of a general purpose index of development.

Appendix A3.1 Summary of study findings included in the rating for the behavior domain[1]

Outcome measure	Study sample	Sample size (classrooms/ students)	Author's findings from the study: Mean outcome (standard deviation[2]) — *Facing History and Ourselves* group[3] (column 1)	Author's findings from the study: Mean outcome (standard deviation[2]) — Comparison group (column 2)	WWC calculations: Mean difference[4] (column 1– column 2)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Shultz, Barr, & Selman, 2001 (quasi-experimental design)								
Self-reported fighting	Grade 8	22/346	1.64 (2.41)	2.24 (3.75)	0.60	0.20	ns	+8
Domain average[8] for behavior						0.20	ns	+8

ns = not statistically significant

1. This appendix reports findings considered for the effectiveness rating and the improvement index.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. The *Facing History and Ourselves* group mean equals the comparison group mean (column 2) plus the mean difference (column 3). The computation of the mean difference took into account the pretest difference between the study groups.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group. The measure, fighting, was reversed so that a positive difference would favor the intervention group.
5. For an explanation of the effect size calculation, please see the Technical Details of WWC-Conducted Computations.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups. These significance levels differ from those in the original study paper but are based on information provided to the WWC by the study author as an amendment to the study report. The level of statistical significance was calculated by the WWC and corrects for clustering within classrooms or schools and for multiple comparisons. For an explanation see the WWC Tutorial on Mismatch. See the Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of *Facing History and Ourselves* a correction for clustering was needed, so the significance levels differ from those reported in the original study.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting favorable results.
8. This row provides the study average, which is also the domain average in this case. The WWC-computed domain average effect size is a simple average rounded to two decimal places. The domain improvement index is calculated from the average effect size.

Appendix A3.2 Summary of study findings included in the rating for the knowledge, attitudes, and values domain[1]

Outcome measure	Study sample[3]	Sample size (classrooms/ students)	Author's findings from the study: Mean outcome (standard deviation[2]) — *Facing History and Ourselves* group[4] (column 1)	Author's findings from the study: Mean outcome (standard deviation[2]) — Comparison group (column 2)	WWC calculations: Mean difference[5] (column 1– column 2)	WWC calculations: Effect size[6]	WWC calculations: Statistical significance[7] (at $\alpha = 0.05$)	WWC calculations: Improvement index[8]
Shultz, Barr, & Selman, 2001 (quasi-experimental design)								
Relationship maturity (best response score)	Grade 8	22/346	2.22 (0.30)	2.07 (0.38)	0.15	0.45	ns	+17
Relationship maturity (response rating score)	Grade 8	22/346	2.07 (0.16)	2.03 (0.19)	0.04	0.23	ns	+9
Racism	Grade 8	22/346	3.29 (0.45)	3.17 (0.45)	0.12	0.27	ns	+10
Ethnic identity	Grade 8	22/346	3.48 (0.78)	3.60 (0.30)	–0.12	-0.19	ns	–7
Civic attitudes and participation	Grade 8	22/346	2.99 (0.51)	2.90 (0.60)	0.09	0.16	ns	+7
Moral reasoning (P score)	Grade 8	9/211	23.00 (12.50)	24.20 (12.60)	–1.20	-0.10	ns	–4
Moral reasoning (D score)	Grade 8	9/211	15.60 (4.30)	16.10 (9.20)	–0.50	-0.07	ns	–3
Domain average[9] for knowledge, attitudes, and values						0.11	ns	+4

ns = not statistically significant

1. This appendix reports summary findings considered for the effectiveness rating and the improvement index.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. The *Facing History and Ourselves* group mean equals the comparison group mean (column 2) plus the mean difference (column 3). The computation of the mean difference took into account the pretest difference between the study groups.
4. The *Facing History and Ourselves* mean equals the comparison group mean (column 2) plus the mean difference (column 3). The mean difference reflects the mean difference that takes into account change from baseline that was used for the effect size calculation.
5. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
6. For an explanation of the effect size calculation, please see the Technical Details of WWC-Conducted Computations.
7. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups. These significance levels differ from those in the original study paper but are based on information provided to the WWC by the study author as an amendment to the study report. The level of statistical significance was calculated by the WWC and corrects for clustering within classrooms or schools and for multiple comparisons. For an explanation see the WWC Tutorial on Mismatch. See the Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of *Facing History and Ourselves* corrections for clustering and multiple comparisons were needed, so the significance levels differ from those reported in the original study.
8. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting favorable results.
9. This row provides the study average, which is also the domain average in this case. The WWC-computed domain average effect size is a simple average rounded to two decimal places. The domain improvement index is calculated from the average effect size.

Appendix A4.1 Rating for the behavior domain

The WWC rates an intervention's effects for a given outcome domain as positive, potentially positive, mixed, no discernible effects, potentially negative, or negative.[1]

For the outcome domain of behavior, the WWC rated *Facing History and Ourselves* as having no discernible effects. It did not meet the criteria for positive effects because it only had one study. In addition, it did not meet the criteria for other ratings (potentially positive effects, mixed effects, potentially negative effects, and negative effects) because the single study that met WWC standards did not show statistically significant or substantively important effects.

Rating received

No discernible effects: No affirmative evidence of effects.

- Criterion 1: None of the studies shows a statistically significant or substantively important effect, either positive or negative.

  Met. The WWC analysis found no statistically significant or substantively important effects in this domain.

Other ratings considered

Positive effects: Strong evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: Two or more studies showing statistically significant *positive* effects, at least one of which met WWC evidence standards for a strong design.

  Not met. The WWC analysis found no statistically significant positive effects in this domain. *Facing History and Ourselves* had only one evaluation study meeting WWC evidence standards that reported findings on behavior, and so did not meet this criterion. Further, that study did not meet WWC evidence standards for a strong design, because it used a QED rather than an RCT design.

- Criterion 2: No studies showing statistically significant or substantively important *negative* effects.

  Met. The WWC analysis found no statistically significant or substantively important negative effects in this domain.

Potentially positive effects: Evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect, thus qualifying as a *positive* effect.

  Not met. The WWC analysis found no statistically significant or substantively important positive effects in this domain.

- Criterion 2: No studies showing a statistically significant or substantively important negative effect. Fewer or the same number of studies showing *indeterminate* effects than showing statistically significant or substantively important *positive* effects.

  Not met. No studies showed a statistically significant or substantively important negative effect. Because one study showed indeterminate effects and no studies showed statistically significant or substantively important positive effects, *Facing History and Ourselves* did not meet this criterion.

(continued)

Appendix A4.1 Rating for the behavior domain *(continued)*

Mixed effects: Evidence of both positive and negative effects.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect. At least one study showing a statistically significant or substantively important *negative* effect, but no more such studies than the number showing a statistically significant or substantively important *positive* effect

  Not met. The WWC analysis found no statistically significant or substantively important positive or negative effects in this domain.

OR

- Criterion 2: At least one study showing a statistically significant or substantively important effect, and more studies showing an *indeterminate* effect than showing a statistically significant or substantively important effect.

  Not met. The WWC analysis found no statistically significant or substantively important effects in this domain.

Potentially negative effects: Evidence of a negative effect with no overriding contrary evidence.

- Criterion 1: At least one study showing a statistically significant or substantively important *negative* effect.

  Not met. The WWC analysis found no statistically significant or substantively important negative effects in this domain.

- Criterion 2: No studies showing a statistically significant or substantively important *positive* effect, or more studies showing statistically significant or substantively important *negative* effects than showing statistically significant or substantively important *positive* effects.

  Met. The WWC analysis found no statistically significant or substantively important positive effects in this domain.

Negative effects: Strong evidence of a negative effect with no overriding contrary evidence.

- Criterion 1: Two or more studies showing statistically significant *negative* effects, at least one of which is based on a strong design.

  Not met. The WWC analysis found no statistically significant negative effects in this domain. *Facing History and Ourselves* had only one evaluation study meeting WWC evidence standards that reported findings on behavior, and so did not meet this criterion. Further, that study did not meet WWC evidence standards for a strong design, because it used a QED rather than an RCT design.

- Criterion 2: No studies showing statistically significant or substantively important *positive* effects.

  Met. The WWC analysis found no statistically significant or substantively important positive effects in this domain.

1. For rating purposes, the WWC considers the statistical significance of individual outcomes and the domain level effect. The WWC also considers the size of the domain level effect for ratings of potentially positive effects. See the WWC Intervention Rating Scheme for a complete description.

Appendix A4.2 Rating for the knowledge, attitudes, and values domain

The WWC rates an intervention's effects for a given outcome domain as positive, potentially positive, mixed, no discernible effects, potentially negative, or negative.[1]

For the outcome domain of knowledge, attitudes, and values, the WWC rated *Facing History and Ourselves* as having no discernible effects. It did not meet the criteria for positive effects because it only had one study. In addition, it did not meet the criteria for other ratings (potentially positive effects, mixed effects, potentially negative effects, and negative effects) because the single study that met WWC standards did not show statistically significant or substantively important effects.

Rating received

No discernible effects: No affirmative evidence of effects.

- Criterion 1: None of the studies shows a statistically significant or substantively important effect, either positive or negative.

  Met. The WWC analysis found no statistically significant or substantively important effects in this domain.

Other ratings considered

Positive effects: Strong evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: Two or more studies showing statistically significant *positive* effects, at least one of which met WWC evidence standards for a strong design.

  Not met. The WWC analysis found no statistically significant positive effects in this domain. *Facing History and Ourselves* had only one evaluation study meeting WWC evidence standards that reported findings on knowledge, attitudes, and values, and so did not meet this criterion. Further, that study did not meet WWC evidence standards for a strong design, because it used a QED rather than an RCT design.

- Criterion 2: No studies showing statistically significant or substantively important *negative* effects.

  Met. The WWC analysis found no statistically significant or substantively important negative effects in this domain.

Potentially positive effects: Evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect, thus qualifying as a *positive* effect.

  Not met. The WWC analysis found no statistically significant or substantively important positive effects in this domain.

- Criterion 2: No studies showing a statistically significant or substantively important *negative* effect. Fewer or the same number of studies showing *indeterminate* effects than showing statistically significant or substantively important *positive* effects.

  Not met. No studies showed a statistically significant or substantively important negative effect. Because one study showed indeterminate effects and no studies showed statistically significant or substantively important positive effects, *Facing History and Ourselves* did not meet this criterion.

(continued)

Appendix A4.2 Rating for the knowledge, attitudes, and values domain *(continued)*

Mixed effects: Evidence of both positive and negative effects.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect. At least one study showing a statistically significant or substantively important *negative* effect, but no more such studies than the number showing a statistically significant or substantively important *positive* effect.

  Not met. The WWC analysis found no statistically significant or substantively important positive or negative effects in this domain.

OR

- Criterion 2: At least one study showing a statistically significant or substantively important effect, and more studies showing an indeterminate effect than showing a statistically significant or substantively important effect.

  Not met. The WWC analysis found no statistically significant or substantively important effects in this domain.

Potentially negative effects: Evidence of a negative effect with no overriding contrary evidence.

- Criterion 1: At least one study showing a statistically significant or substantively important *negative* effect.

  Not met. The WWC analysis found no statistically significant or substantively important negative effects in this domain.

- Criterion 2: No studies showing a statistically significant or substantively important *positive* effect, or more studies showing statistically significant or substantively important *negative* effects than showing statistically significant or substantively important *positive* effects.

  Met. The WWC analysis found no statistically significant or substantively important positive effects in this domain.

Negative effects: Strong evidence of a negative effect with no overriding contrary evidence.

- Criterion 1: Two or more studies showing statistically significant *negative* effects, at least one of which is based on a strong design.

  Not met. The WWC analysis found no statistically significant negative effects in this domain. *Facing History and Ourselves* had only one evaluation study meeting WWC evidence standards that reported findings on knowledge, attitudes, and values, and so did not meet this criterion. Further, that study did not meet WWC evidence standards for a strong design, because it used a QED rather than an RCT design.

- Criterion 2: No studies showing statistically significant or substantively important *positive* effects.

  Met. The WWC analysis found no statistically significant positive effects in this domain.

1. For rating purposes, the WWC considers the statistical significance of individual outcomes and the domain level effect. The WWC also considers the size of the domain level effect for ratings of potentially positive effects. See the WWC Intervention Rating Scheme for a complete description.