Teacher Observations of Preschoolers’ Social-Emotional Behavior:

A Formative Evaluation

BY

CLAIRE CHRISTENSEN

B.S., Illinois State University, 2008

M.A., University of Illinois at Chicago, 2011

DOCTORAL DISSERTATION

Submitted as partial fulfillment of the requirements

for the degree of Doctor of Philosophy in Psychology

in the Graduate College of the

University of Illinois at Chicago, 2015

Chicago, Illinois

Dissertation Committee:

Kate Zinsser, Chair and Advisor
Carol Myford, Educational Psychology
Ed Trickett, Psychology
Amanda Roy, Psychology
Celene Domitrovich, Collaborative for Academic Social and Emotional Learning
To my fiancé, Casey. You see the best in me as a researcher and a person, and I have spent my graduate career working toward that ideal. Your thinking has sharpened my work, your edits have clarified my writing, and your support has strengthened my confidence.

To my mother, Sherry. Thank you for sharing your love of words and of learning. Thank you for showing me that even the hardest work can be completed. Thank you for sharing the love and memory of those who are no longer with us: Jimmy Vannucci Jr., Jeanette Vannucci, Jimmy Vannucci Sr., and Kathryn Vannucci.

To my father, David. Thank you for being a role model of optimism, resilience, and persistence. Thank you for fostering my early love of math and science. Thank you for moving me nine times throughout my higher education. Thank you for all the cake.

To my brother, Ross. Whether I succeed or fail I know you will love me just the same, and that has given me such comfort. Thank you for making me laugh and convincing me to have fun. You are the best brother anyone could ask for.

To my advisor, Dr. Kate Zinsser. Working with you has been the best part of my graduate education. You have been an invaluable role model of productivity and positivity. In your lab I came to feel like a success, and I will be forever thankful for that.

To Dr. Carol Myford. Thank you for sharing your passion for measurement and evaluation. Your teaching approach made this work seem accessible and fascinating. Your mentorship has shaped my interests, and I am excited to see where they lead.

To my cousin, Brandon. Early in my life you showed me the magic of feeling passionate about my work. All of my best accomplishments since, including this degree, have been in pursuit of the wonder and delight I felt making movies with you as a child. I will always be grateful to know the joy you taught me.
ACKNOWLEDGEMENTS

I would like to thank my dissertation committee—Drs. Kate Zinsser, Carol Myford, Ed Trickett, Amanda Roy, and Celene Domitrovich—for their support and guidance. I would also like to thank the administrators and staff at the University of Illinois at Chicago Children’s Center for their collaboration. Finally I owe many thanks to the undergraduate research assistants in the Social Emotional Teaching and Learning Laboratory, especially to Luz Torres, Madlen Zamfirova, and Phil George.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.  1. INTRODUCTION ...........................................</td>
<td>1</td>
</tr>
<tr>
<td>A. Social-Emotional Competencies ......................</td>
<td>1</td>
</tr>
<tr>
<td>1. Emotion Expression ....................................</td>
<td>1</td>
</tr>
<tr>
<td>2. Involvement ..............................................</td>
<td>2</td>
</tr>
<tr>
<td>3. Emotion Regulation .....................................</td>
<td>2</td>
</tr>
<tr>
<td>4. Relationship Skills ....................................</td>
<td>3</td>
</tr>
<tr>
<td>5. Need for Assessment ....................................</td>
<td>3</td>
</tr>
<tr>
<td>B. Live Observations for Formative Assessment ........</td>
<td>4</td>
</tr>
<tr>
<td>C. Minnesota Preschool Affect Checklist ..............</td>
<td>6</td>
</tr>
<tr>
<td>1. Minnesota Preschool Affect Checklist—Teacher ...</td>
<td>10</td>
</tr>
<tr>
<td>D. Teacher Observations ..................................</td>
<td>10</td>
</tr>
<tr>
<td>E. Rater Training ...........................................</td>
<td>11</td>
</tr>
<tr>
<td>1. Frame-of-Reference Training .........................</td>
<td>12</td>
</tr>
<tr>
<td>2. Computerized Rater Training ..........................</td>
<td>13</td>
</tr>
<tr>
<td>F. The Present Study .......................................</td>
<td>13</td>
</tr>
<tr>
<td>1. Aim 1: Assess the Efficacy of the MPAC-T Rater Training Program</td>
<td>14</td>
</tr>
<tr>
<td>2. Aim 2: Identify Areas for Improvement in the MPAC-T Rater Training Program</td>
<td>14</td>
</tr>
<tr>
<td>3. Aim 3: Describe Factors that Could Affect the Accuracy of Teachers’ MPAC-T Ratings</td>
<td>15</td>
</tr>
<tr>
<td>II. 2. METHOD ..................................................</td>
<td>16</td>
</tr>
<tr>
<td>A. Computerized MPAC-T Rater Training Program .. .</td>
<td>16</td>
</tr>
<tr>
<td>B. Participants ..............................................</td>
<td>17</td>
</tr>
<tr>
<td>1. Recruitment .............................................</td>
<td>17</td>
</tr>
<tr>
<td>2. Screening ..................................................</td>
<td>18</td>
</tr>
<tr>
<td>3. Participant Characteristics ...........................</td>
<td>18</td>
</tr>
<tr>
<td>C. Measure .....................................................</td>
<td>19</td>
</tr>
<tr>
<td>D. Videos ......................................................</td>
<td>21</td>
</tr>
<tr>
<td>1. Training Videos .........................................</td>
<td>22</td>
</tr>
<tr>
<td>2. Assessment Videos ......................................</td>
<td>24</td>
</tr>
<tr>
<td>E. Expert Ratings ...........................................</td>
<td>24</td>
</tr>
</tbody>
</table>
# TABLE OF CONTENTS (continued)

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>F. Procedures</td>
<td>24</td>
</tr>
<tr>
<td>1. Pre-Test</td>
<td>25</td>
</tr>
<tr>
<td>2. Laboratory Session</td>
<td>25</td>
</tr>
<tr>
<td>3. Follow-Up</td>
<td>26</td>
</tr>
<tr>
<td>G. Mixed-Methods Research Design</td>
<td>27</td>
</tr>
<tr>
<td>H. Focus Group Transcription, Coding, and Reliability</td>
<td>27</td>
</tr>
<tr>
<td>III. 3. RESULTS</td>
<td>30</td>
</tr>
<tr>
<td>A. Completion Rates</td>
<td>30</td>
</tr>
<tr>
<td>B. Aim 1: Assess the Efficacy of the MPAC-T Rater Training Program</td>
<td>30</td>
</tr>
<tr>
<td>1. RQ1: How does Rater Training Influence the Accuracy of Teachers’</td>
<td>30</td>
</tr>
<tr>
<td>MPAC-T Ratings?</td>
<td></td>
</tr>
<tr>
<td>2. RQ2: How Accurate are Teachers’ MPAC-T Ratings after Training?</td>
<td>30</td>
</tr>
<tr>
<td>3. RQ3: How Internally Reliable are Teachers’ MPAC-T Ratings after</td>
<td>31</td>
</tr>
<tr>
<td>Training?</td>
<td></td>
</tr>
<tr>
<td>4. RQ4: How Stable are Teachers’ MPAC-T Ratings One Week after Training?</td>
<td>36</td>
</tr>
<tr>
<td>5. Aim 1 Summary</td>
<td>37</td>
</tr>
<tr>
<td>C. Aim 2: Identify Areas for Improvement in the MPAC-T Rater Training</td>
<td>37</td>
</tr>
<tr>
<td>Program</td>
<td></td>
</tr>
<tr>
<td>1. RQ5: On Which MPAC-T Items are Teachers’ Ratings Least Accurate?</td>
<td>38</td>
</tr>
<tr>
<td>2. RQ6: On which MPAC-T videos are teachers’ ratings least accurate?</td>
<td>39</td>
</tr>
<tr>
<td>3. RQ7: What Difficulties Did Teachers Encounter When Rating Videos?</td>
<td>43</td>
</tr>
<tr>
<td>4. RQ8: What Difficulties Did Teachers Encounter When Using the Training Program?</td>
<td>47</td>
</tr>
<tr>
<td>5. Aim 2 Summary</td>
<td>48</td>
</tr>
<tr>
<td>D. Aim 3: Describe Factors that Could Affect the Accuracy of Teachers’</td>
<td>48</td>
</tr>
<tr>
<td>MPAC-T Ratings</td>
<td></td>
</tr>
<tr>
<td>1. RQ9: What Teacher Characteristics are Associated with the Highest</td>
<td>48</td>
</tr>
<tr>
<td>Accuracy?</td>
<td></td>
</tr>
<tr>
<td>2. RQ10: What Factors Might Affect Teachers’ MPAC-T Ratings in the</td>
<td>49</td>
</tr>
<tr>
<td>Classroom?</td>
<td></td>
</tr>
<tr>
<td>3. Aim 3 Summary</td>
<td>51</td>
</tr>
<tr>
<td>IV. 4. DISCUSSION</td>
<td>52</td>
</tr>
<tr>
<td>A. Efficacy of the MPAC-T Rater Training Program</td>
<td>52</td>
</tr>
<tr>
<td>1. Trained Teachers Can Use the MPAC-T Accurately</td>
<td>52</td>
</tr>
<tr>
<td>2. Teachers’ MPAC-T Ratings are Reliable</td>
<td>53</td>
</tr>
<tr>
<td>B. Areas for Improvement in the MPAC-T Rater Training Program</td>
<td>54</td>
</tr>
</tbody>
</table>
# TABLE OF CONTENTS (continued)

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Provide More Detailed Information about Items</td>
<td>54</td>
</tr>
<tr>
<td>2. Provide More Focused Guidance on Items that Were Especially Difficult for Teachers to Rate Accurately</td>
<td>55</td>
</tr>
<tr>
<td>3. Shorten the Reliability Testing Phase</td>
<td>57</td>
</tr>
<tr>
<td>4. Correct Technical Glitches in the MPAC-T Rater Training Program</td>
<td>57</td>
</tr>
<tr>
<td>C. Factors that Could Affect the Accuracy of Teachers’ MPAC-T Ratings</td>
<td>58</td>
</tr>
<tr>
<td>1. Assistant versus Lead Teachers</td>
<td>58</td>
</tr>
<tr>
<td>2. Knowledge of the Child</td>
<td>58</td>
</tr>
<tr>
<td>D. Future Directions</td>
<td>59</td>
</tr>
<tr>
<td>E. Limitations</td>
<td>60</td>
</tr>
<tr>
<td>F. Conclusion</td>
<td>62</td>
</tr>
<tr>
<td>CITED LITERATURE</td>
<td>63</td>
</tr>
<tr>
<td>APPENDICES</td>
<td>Error! Bookmark not defined.</td>
</tr>
<tr>
<td>Appendix A</td>
<td>73</td>
</tr>
<tr>
<td>Appendix B</td>
<td>74</td>
</tr>
<tr>
<td>Appendix C</td>
<td>75</td>
</tr>
<tr>
<td>Appendix D</td>
<td>80</td>
</tr>
<tr>
<td>Appendix E</td>
<td>83</td>
</tr>
<tr>
<td>VITA</td>
<td>85</td>
</tr>
</tbody>
</table>
LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. COMPARISON OF ITEMS ON THE MINNESOTA PRESCHOOL AFFECT CHECKLIST-TEACHER (MPAC-T) AND MPAC-REVISED/SHORTENED</td>
<td>8</td>
</tr>
<tr>
<td>II. PARTICIPANT CHARACTERISTICS</td>
<td>20</td>
</tr>
<tr>
<td>III. ASSESSMENT VIDEO CHARACTERISTICS AND PARTICIPANT ACCURACY ON THE MINNESOTA PRESCHOOL AFFECT CHECKLIST-TEACHER</td>
<td>23</td>
</tr>
<tr>
<td>IV. TEACHER FOCUS GROUP CODES AND THEMES</td>
<td>29</td>
</tr>
<tr>
<td>V. DISTRIBUTION OF TEACHER KAPPA VALUES</td>
<td>33</td>
</tr>
<tr>
<td>VI. TABLE VI</td>
<td>34</td>
</tr>
<tr>
<td>VII. SUMMARY STATISTICS FOR MULTI-FACET RASCH MEASUREMENT ANALYSES OF MPAC-T INTRA-RATER RELIABILITY AND ACCURACY</td>
<td>34</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Variable map.</td>
<td>41</td>
</tr>
<tr>
<td>2. Rates of under- and over-rating for MPAC-T items and videos.</td>
<td>42</td>
</tr>
</tbody>
</table>
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOR</td>
<td>Frame of Reference</td>
</tr>
<tr>
<td>MFRM</td>
<td>Multi-Facet Rasch Measurement</td>
</tr>
<tr>
<td>MPAC</td>
<td>Minnesota Preschool Affect Checklist</td>
</tr>
<tr>
<td>MPAC-R</td>
<td>Minnesota Preschool Affect Checklist—Revised</td>
</tr>
<tr>
<td>MPAC-R/S</td>
<td>Minnesota Preschool Affect Checklist—Revised/Shortened</td>
</tr>
<tr>
<td>MPAC-T</td>
<td>Minnesota Preschool Affect Checklist—Teacher</td>
</tr>
<tr>
<td>RQ</td>
<td>Research Question</td>
</tr>
</tbody>
</table>
SUMMARY

Social-emotional competencies—skills for managing one’s emotions and relationships—are integral to children’s academic and social success. Early-childhood educators need quality measures to track and support these competencies. One such measure is the Minnesota Preschool Affect Checklist-Teacher (MPAC-T), a new teacher observation tool. This paper describes a formative evaluation of a rater training program for teachers learning to use the MPAC-T.

Twenty-two early-childhood educators participated in a pretest; a laboratory visit including MPAC-T training, a post-test, and focus groups; and a one-week follow-up. Teachers rated videos more accurately at post-test than at pre-test. Teachers’ post-test ratings were moderately accurate and consistent over time. Multi-facet Rasch measurement analyses revealed specific items and videos that were rated least accurately. Focus group responses highlighted specific ways the MPAC-T training program can be revised to improve the accuracy of teachers’ ratings.
1. INTRODUCTION

Social-emotional competencies—skills for managing one’s relationships and emotions—are integral to children’s academic and social development (Weissberg & Cascarino, 2013; Zinsser, Weissberg, & Dusenbury, 2013). Researchers have called for the development of formative assessments that allow teachers to measure and track children’s social-emotional competencies (Bridgeland, Bruce, & Harihan, 2013). One promising new option is the Minnesota Preschool Affect Checklist-Teacher, an early-childhood observation measure. The present study is a formative evaluation of the MPAC-T rater training program for teachers. It employs mixed methods to investigate the training program’s efficacy, areas for improvement, and factors that may affect teachers’ ratings. Findings will be used to improve the MPAC-T rater training program and ultimately support teachers’ assessment of children’s social-emotional competencies.

A. Social-Emotional Competencies

Social-emotional competencies include appropriate emotion expression, involvement in age-appropriate activities, emotion regulation, and relationship skills (Denham et al., 2012a). These competencies are associated with positive social and academic outcomes throughout childhood (Zins, Bloodworth, Weissberg, & Walberg, 2007). Development of social-emotional competencies in early childhood provides a foundation for kindergarten readiness, promoting positive social interaction, classroom engagement, and academic achievement (Bodrova & Leong, 2006; Denham & Brown, 2010).

1. Emotion Expression

Preschoolers’ expression of positive and negative emotions influences their status in the classroom. Children who express more positive than negative affect are perceived as friendlier,
more assertive, more prosocial, more likable, and less aggressive and sad (Denham, 1986; Denham, Renwick, & Holt, 1991; Eisenberg et al., 1995; Sroufe, Schork, Motti, Lawroski, & LaFreniere, 1984; Strayer, 1980). This in turn may affect how peers and instructors interact with students, with more emotionally negative students experiencing lower-quality interactions.

2. **Involvement**

Children are more likely to benefit from the educational environment if they become actively engaged in it. For example, teacher-reported attention and persistence in preschool is positively associated with academic school readiness (Vitiello, Greenfield, Munis, & George, 2011) and with both expressive and receptive vocabulary (Fantuzzo, Perry, & McDermott, 2004). Early classroom involvement predicts later academic success; kindergarten approaches to learning (i.e., persistence, emotion regulation, and attentiveness) positively predict math and reading scores through fifth grade (Li-Grining, Votruba-Drzal, Maldonado-Carreno, & Haas, 2010).

3. **Emotion Regulation**

Students’ reactions to frustration, such as aggression vs. productive discussion, predict academic engagement and success. Emotion regulation predicts concurrent and future classroom engagement and attention to academic tasks (Humphries et al., 2012; Shields et al., 2001; Trentacosta & Izard, 2007). This engagement translates to success. Child emotion regulation is associated with teacher reports of academic success and classroom productivity (Graziano, Reavis, Keane, & Calkins, 2007). It is also associated with performance on standardized literacy and mathematics assessments (Graziano et al., 2007; McClelland et al., 2007). This association is lasting; preschool emotion regulation positively predicts kindergarten academic success (Howse, Calkins, Anastopoulos, & Shelton, 2003).
4. Relationship Skills

Relationship skills include leading and joining groups, empathizing with others, and behaving in a prosocial manner. These skills are positively associated with peer and adult social interactions (Humphries, Keenan, Wakschlag, 2012), and they predict future social status (Ladd, Price, & Hart, 1990). Children with stronger peer skills are also more academically successful. Positive interactive play behavior in preschool is associated with greater engagement in classroom learning activities (Coolahan, Fantuzzo, Mendez, & McDermott, 2000). Early prosocialness further predicts later academic achievement, and it is a better predictor than early academic achievement (Caprara, Barbaranelli, Pastorelli, Bandura, & Zimbardo 2000).

5. Need for Assessment

Acknowledging the impact of early social-emotional competencies on student outcomes, policymakers have established preschool social-emotional learning standards in all 50 of the United States (Dusenbury et al., 2015). To meet these standards, teachers need practical, quality measures for assessing and tracking children’s social-emotional competencies (Zinsser et al., 2013). Although teachers have many sophisticated ways to measure children’s academic performance, there are fewer high-quality measures of children’s social-emotional competencies (Chafouleas et al., 2010; Riley-Tillman, Kalberer, & Chafouleas, 2005; Snow & Van Hemel, 2008). Early-childhood educators’ options are further limited by young children’s inability to easily and accurately report on their own emotions. Early-childhood educators must therefore use alternative assessment methods, such as observations (Denham, Wyatt, Bassett, Echeverria, & Knox, 2009). Conducting regular systematic observations may help early-childhood educators track students’ progress toward social-emotional learning standards and support their growth.
B. **Live Observations for Formative Assessment**

Formative assessment is particularly well-suited for supporting social-emotional instruction (Bridgeland et al., 2013). Formative assessments are ongoing measurements of students’ strengths and weaknesses which can be used to guide instructional decisions (Denham, 2006). They are distinct from summative assessments—one-time performance measures used to index achievement. Because formative assessment provides rapid, usable feedback to instructors, it is well-aligned with recommendations of the National Research Council (Snow & Van Hemel, 2008), the National Association for the Education of Young Children (NAEYC, 2009), and the Collaborative for Academic, Social, and Emotional Learning (Kendziora, Weissberg, & Dusenbury, 2011).

Most measures of children’s social-emotional competencies rely on retrospective teacher or parent reports, which are more appropriate for clinical screening than for formative assessment (Denham, Ji, & Hamre, 2010; Volpe & Chafouleas, 2011; Webster-Stratton & Lindsay, 1999). Commonly used retrospective report measures of children’s social-emotional competencies include the Rothbart Child Behavior Questionnaire (Cole, 1991), the Devereux Early Childhood Assessment (LeBuffe & Naglieri, 1999), and the Emotion Regulation Checklist (Shields et al., 2001). These measures ask parents or teachers about a child’s past tendencies over a period of weeks to months (Gladman & Lancaster, 2003; Ramsay, Reynolds, & Kamphaus, 2002). For example, the Rothbart Child Behavior Questionnaire asks about the child’s tendencies over the past six months (Denham et al., 2010). Retrospective reports therefore yield information about a child’s general traits (Gladman & Lancaster, 2003; Ramsay, Reynolds, & Kamphaus, 2002), but are less appropriate for measuring behavioral changes over a short timespan.
To provide real-time feedback on instructional practices, formative assessments should reflect immediate behavior, rather than past tendencies. Live observational measures—records of a child’s behavior during a short timespan—may be more appropriate for this purpose. They reflect behavior in the moment, rather than past tendencies or traits. Further, live observations can be conducted frequently to measure change. Retrospective reports, meanwhile, are insensitive to changes occurring over a short timespan (Riley-Tillman et al., 2005) and often too lengthy for teachers to complete frequently (Briesch, Chafouleas, & Riley-Tillman, 2010).

Because observational measures require the rater to record behavior as it occurs, they yield more accurate information (Riley-Tillman et al., 2005), whereas retrospective reports tend to be more subjective (McConaughy, 1993; Cost & Simpson, 2004) and prone to rater effects (Abikoff, Courtney, Pelham, & Koplewicz, 1993; Briesch et al., 2010; Cohen & Kasen, 1999). Indeed, scores on observational measures and retrospective reports are only weakly to moderately correlated (Cost & Simpson, 2004). Of the two, direct observations of child behavior are better predictors of future adjustment (Patterson & Forgatch, 1995). For these reasons NAEYC recommends the use of observations for early-childhood formative assessments (NAEYC, 2003), and researchers have called for the development of teacher observation measures that are valid, reliable and accompanied by rater training (Hintze, Volpe, & Shapiro, 2002).
C. **Minnesota Preschool Affect Checklist**

The Minnesota Preschool Affect Checklist (MPAC) is an observational measure of young children’s social-emotional competencies—specifically expressed emotions, reactions to frustration, involvement in classroom activities, and interaction with others (Denham et al., 2012a). To use the MPAC-T, trained observers rate a child’s behavior as it occurs during five minutes of free play. Observers repeat this process for a given child four times over four consecutive weeks.

Since its development over three decades ago, the MPAC has undergone revisions while retaining its essential characteristics as an observational measure of children’s social-emotional competencies. The MPAC was originally developed by Sroufe and colleagues (1984) as a 52-item clinical screening tool. Denham and colleagues then adapted it into a 62-item observational measure for research, the MPAC-R (Denham & Burton, 1996; Denham, Zahn-Waxler, Cummings, & Ianotti, 1991). Later, Denham and colleagues shortened the MPAC-R to an 18-item measure, the MPAC-R/S, while gathering additional evidence to support the reliability and validity of inferences made from scores on the revised instrument (Denham et al., 2012a). The MPAC-R/S, included in Table 1, allows researchers to assess three domains of children’s behavior: emotionally positive/productive, emotionally negative/aggressive, and emotionally regulated/prosocial behavior.

When employed for research purposes, the MPAC-R and R/S have demonstrated reliability and validity. They have exhibited interrater reliability (Denham & Burton, 1996; Denham et al., 1991; Denham, Bassett, Zinsser, & Wyatt, 2014) and internal consistency (Denham et al., 2012a). As evidence of concurrent validity, high scores on these instruments are associated with preschool success, maternal affect, and emotion knowledge (Denham et al.,
The MPAC-R/S further exhibits predictive validity; preschool MPAC-R/S ratings of emotionally negative/aggressive behaviors are associated with poorer kindergarten school adjustment and lower academic success (Denham et al., 2012a).

The MPAC is well suited for formative assessment. Administration requires only five minutes per child, allowing for more frequent use. Further, it is sensitive to change; the MPAC has detected both age and pre-post intervention changes (Denham & Burton, 1996; Denham et al., 1991). The MPAC-R/S is also sensitive to instructional practice; children’s MPAC-R/S
<table>
<thead>
<tr>
<th>MPAC-T Item</th>
<th>MPAC-R/S Item</th>
<th>Reverse Scored?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Showed positive emotions in any manner (i.e., facial, vocal, or bodily) that made sense given the situation.</td>
<td>1. The child displays positive affect in any manner (i.e., facial, vocal, or bodily affect). The child’s behaviors must match the context of a given situation.</td>
<td></td>
</tr>
<tr>
<td>1a. The positive emotion was directed a particular person.</td>
<td>2. The child directs positive affect specifically at a particular person when already in contact with them. Affect is directed at a specific person. This behavior does not have to be in response to someone, but someone must be the target of the child’s positive affect.</td>
<td></td>
</tr>
<tr>
<td>1b. The positive emotion was not directed at a particular person.</td>
<td>3. The child displays positive affect when in a social situation but does not direct it to anyone in particular.</td>
<td></td>
</tr>
<tr>
<td>2. Showed negative emotions in any manner (i.e., facial, vocal, or bodily affect) that made sense given the situation.</td>
<td>4. The child displays negative affect in any manner (i.e., facial, vocal, or bodily affect). The child’s behaviors must match the context of a given situation.</td>
<td>Yes</td>
</tr>
<tr>
<td>2a. The negative emotion was directed at a specific person.</td>
<td>5. The child directs negative affect specifically at a particular person when already in contact with them. Affect is directed at a specific person. This behavior does not have to be in response to someone, but someone must be the target of the child’s negative affect.</td>
<td>Yes</td>
</tr>
<tr>
<td>3. Was engaged and engrossed in an activity that has a positive emotional function.</td>
<td>6. The child is engrossed, absorbed, intensely involved in activity. The child is emotionally invested in creative, productive, thematically organized, or other activity that has a positive emotional function. The child does not seem to notice what is going on around him/her.</td>
<td></td>
</tr>
<tr>
<td>4. Was involved in an activity that he/she helped to organize.</td>
<td>7. The child is involved in an activity that he/she organizes for himself/herself; s/he is independent.</td>
<td></td>
</tr>
<tr>
<td>5. Looked “emotionally absent” and showed no involvement in an activity</td>
<td>8. Vacant: The child displays a very flat, unexpressive, detached face; shows no involvement in an activity; and looks “emotionally absent.”</td>
<td>Yes</td>
</tr>
<tr>
<td>6. Looked fidgety and uninterested or uninvested in the activity but still “tuned in” and aware of his/her surroundings.</td>
<td>9. Listless: The child looks fidgety and uninvested in the activity but still “emotionally present;” the child stays in one area but shows little/no involvement in activities or social interaction.</td>
<td>Yes</td>
</tr>
<tr>
<td>MPAC-T Item</td>
<td>MPAC-R/S Item</td>
<td>Reverse Scored?</td>
</tr>
<tr>
<td>---------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>7. Reacted to a frustrating situation by aggressively attacking another person either verbally or physically.</td>
<td>10. The child displays context-related interpersonal aggression (verbal or physical). Someone does something to which the child responds with aggression (emotionally arousing preceding event must be observed).</td>
<td>Yes</td>
</tr>
<tr>
<td>8. Took frustration out on an object.</td>
<td>11. The child hits, kicks, shoves, knocks over, or throws objects (emotionally arousing preceding event must be observed).</td>
<td>Yes</td>
</tr>
<tr>
<td>9. Reacted to frustration by talking about the problem and did not stay upset.</td>
<td>13. The child promptly verbally expresses feelings arising from a problem situation, then moves on to the same or a new activity (versus withdrawing, displacing the affect onto others or objects, or staying upset).</td>
<td></td>
</tr>
<tr>
<td>9a. In his/her reaction to frustration, the child was neutral or positive.</td>
<td>14. The child shows primarily neutral or positive affect.</td>
<td></td>
</tr>
<tr>
<td>10. Physically attacked another child for no apparent reason.</td>
<td>12. The child displays unprovoked physical interpersonal aggression with no preceding provocation by the victim.</td>
<td>Yes</td>
</tr>
<tr>
<td>11. Smoothly approached an ongoing activity and easily got actively involved.</td>
<td>15. The child smoothly approaches an already ongoing activity and gets actively involved. The child does not disrupt or antagonize other children as he/she approaches the activity.</td>
<td></td>
</tr>
<tr>
<td>12. The child displayed leadership skills by organizing an activity in which another child or children “followed the lead” and participated.</td>
<td>12. Successful leadership: The child plays an organizing role in an activity in which another child or children “follow the lead” and participate.</td>
<td></td>
</tr>
<tr>
<td>13. Took turns with classroom materials.</td>
<td>16. Taking turns: The child plays with a toy or participates in an activity and then allows another to do the same. A clear beginning and end of each child’s turn during an activity must be observed.</td>
<td></td>
</tr>
<tr>
<td>14. Cooperated with peers to achieve a common goal.</td>
<td>17. Cooperating: The child jointly works with a peer or group of peers to achieve a common goal.</td>
<td></td>
</tr>
<tr>
<td>15. Shared toys or other materials.</td>
<td>18. The child shares toys or other materials (e.g., crayon, pencil, play dough, etc.).</td>
<td></td>
</tr>
<tr>
<td>16. Engaged in no social interactions for 3 minutes or more.</td>
<td>10. The child engages in no social interaction continuously for 3 minutes or more.</td>
<td>Yes</td>
</tr>
</tbody>
</table>

*Note. Rxn = reaction. *a Item not included on the MPAC-R/S but included on the MPAC-R.*
scores are associated with their teachers’ emotion socialization beliefs and practices (Morris, Denham, Bassett, & Curby, 2013). For these reasons Denham and colleagues (2012b) have stated that a classroom version of the MPAC-R/S may be useful for formative assessment in early-childhood education.

1. Minnesota Preschool Affect Checklist—Teacher

To date the MPAC has only been used by extensively trained researchers. The present study aims to demonstrate that with training, teachers can use an adaption of the MPAC-R/S—the MPAC-Teacher (MPAC-T)—accurately and reliably. The MPAC-T maintains all items from the MPAC-R/S. It also includes two additional items from the MPAC-R, which the test developers identified for inclusion in future version of the MPAC due to their conceptual relevance (Denham, Bassett, & Zinsser, 2012b). In contrast to prior versions, the MPAC-T uses simplified wording to facilitate teacher use with minimal training. See Table 1 for a comparison of MPAC-T and MPAC-R/S items.

The MPAC-T is intended for use as a formative assessment tool in early-childhood education classrooms. Teachers would use it to identify strengths and weaknesses in children’s social-emotional competencies and modify their instructional practices accordingly—for example, to provide more targeted support for areas of weakness (Denham et al., 2012a). However, while prior research (described previously) suggests that the MPAC is appropriate for formative use, the MPAC-T has not been validated for this purpose.

D. Teacher Observations

The MPAC-T enlists teachers as observers of their students’ behavior, which has many advantages. Because teachers are a natural part of the school environment, their observations are less costly and less prone to child reactivity than independent observers’ ratings (Hay, Nelson, &
Hay, 1977). Further, teachers may feel that their own ratings are more valid than others’ ratings (Chafouleas, Riley-Tillman, & Sassu, 2006), and thus may be more likely to base instructional decisions on their own ratings. Indeed, Mathematica Policy Research concluded that it is wiser “to invest in training teachers to be better observers and more reliable assessors than to spend those resources training and paying for outside assessors…” (Mathematica Policy Research, 2007, p. 16).

However, teachers’ ratings can be swayed by irrelevant student, teacher, or environment characteristics. Teachers’ ratings of students may vary with student ethnicity (Pigott & Cowen, 2000; Sbarra & Pianta, 2001; Chang & Sue, 2003), teacher ethnicity (de Ramirez & Shapiro, 2005), or the interaction of both (Downey & Pribesh, 2004; Saft & Pianta, 2001). Additional student characteristics which may bias teachers include socioeconomic status (Phillips & Lonigan, 2010), gender (Jackson & King, 2004), and negative behavior (Phillips & Lonigan, 2010; Riley-Tillman, Christ, Chafouleas, Boice-Mallach, & Briesch, 2011). Teachers’ ratings may also depend on their expectations about children’s behavior (De Los Reyes & Kadzin, 2005), prior notions about a particular child (Tversky & Marsh, 2000), or personal knowledge and skills (Eschen-Leguede, & Peccena, 2006, Brindley, 1998). Teachers’ perceptions of the school environment may also affect their ability to rate children objectively (Pas & Bradshaw, 2013; Kellam, 1994). Therefore teachers must be trained to make objective ratings based on students’ behavior rather than potentially biased inferences.

E. Rater Training

Rater training can improve the accuracy and reliability of teachers’ behavioral observations (Chafouleas, Sanetti, Kilgus, Maggin, 2012; Madle, Neisworth, & Kurtz, 1980). Rater training is the process of introducing raters to scoring criteria and allowing them to work
through examples to foster a common understanding of the measure (Knoch, Read, & Von Randow, 2007). This process improves the reliability and accuracy of raters’ ratings (Chafouleas et al., 2012; Kramer, de Roten, & Drapeau, 2011; LeBel, Kilgus, Briesch, & Chafouleas, 2010; Schlientz, Riley-Tillman, Briesch, Walcott, & Chafouleas, 2009). However, rater training programs may vary in effectiveness depending on their contents and presentation (Schlientz et al., 2009).

1. **Frame-of-Reference Training**

One rater-training approach, frame-of-reference (FOR) training, appears most promising (McIntyre, Smith, & Hassett, 1984). In this approach, raters are trained on a common theory of performance consisting of expert-defined performance dimensions and standards (Bernardin & Buckley, 1981; Woehr & Huffcutt, 1994). Frame-of-reference training involves familiarizing raters with these dimensions and standards, providing examples, and then facilitating rehearsal and feedback (Woehr & Huffcutt, 1994). A meta-analysis found that of several common rater-training approaches, FOR yielded the greatest improvement in rater accuracy (Woehr & Huffcutt, 1994). The efficacy of FOR training was further confirmed in a more recent meta-analysis (Roch, Woehr, Mishra, & Kieszcynska, 2012). Frame-of-reference training may work because it facilitates deeper processing (Athey & McIntyre, 1987; Roch et al., 2012), aids behavioral recall (Woehr, 1994), and facilitates the formation of accurate impressions (Schleicher & Day, 1998; Suls, & Day, 1992).

Frame-of-reference training may be useful for training teachers as behavioral observers. While Bernardin and Buckley (1981) originally developed FOR training to facilitate industrial/organizational performance appraisals, FOR training has since been applied in many relevant contexts, including assessment centers (Goodstone & Lopez, 2001; Schleicher, Day,
Mayes & Riggio, 2002), competency modeling (Lievens & Sanchez, 2007), and therapy (Angkaw, Tran, & Haaga, 2006). Although little research has explored FOR training for teachers, Joyce and Showers (1981) found that teachers learn more from in-service training that includes demonstration and practice with feedback—key components of FOR training.

2. Computerized Rater Training

The MPAC-T makes use of a recent development, computerized rater training. The major advantage of computerized rater training is convenience; trainees can participate remotely when they are available. This allows researchers to conduct large-scale, standardized trainings (Kobak, Engelhard, & Lipsitz, 2006; Kobak, Opler, & Engelhardt, 2007). Computerized training participants report advantages such as convenience, ability to self-pace, and anonymity (Hamilton, Reddel, & Spratt, 2001; Kobak et al., 2007). For these reasons teachers in one study strongly endorsed online rater training for educators (Hamilton et al., 2001). Computerized rater training also supports the use of engaging features such as videos and immediate feedback, which may increase information retention (Kobak et al., 2006). Indeed, computerized rater training is effective; it improves inter-rater reliability (Knoch et al., 2007) and understanding of rating conventions (Kobak et al., 2007, 2011). Thus many large-scale testing organizations have implemented computerized rater training, including Pearson (Wolfe & McVay, 2010) and the Educational Testing Service (Everson & Hines, 2010).

F. The Present Study

The MPAC-T is a new teacher-observation measure of early-childhood social-emotional competencies. I developed a computerized program that trains teachers to rate children using the MPAC-T. The present study is a formative evaluation of the MPAC-T rater training program. Results will be used to improve the program. Early-childhood educators took a pre-test; attended
a laboratory session including rater training, a post-test, and a focus group; and completed
follow-up ratings. This mixed-methods study incorporates quantitative analyses of teachers’
ratings and qualitative analyses of focus group responses. I pursued three aims:

1. **Aim 1: Assess the Efficacy of the MPAC-T Rater Training Program**

   I will describe trained teachers’ MPAC-T ratings using quantitative analyses.

   Specifically, I will pursue four research questions (RQs):

   o **RQ1**: How does rater training influence the accuracy of teachers’ MPAC-T
     ratings?

   o **RQ2**: How accurate are teachers’ MPAC-T ratings after training?

   o **RQ3**: How internally reliable are teachers’ MPAC-T ratings after training?

   o **RQ4**: How stable are teachers’ MPAC-T ratings one week after training?

2. **Aim 2: Identify Areas for Improvement in the MPAC-T Rater Training Program**

   This formative evaluation is part of an iterative process; results will be used to improve the
   MPAC-T rater training program. I will use quantitative and qualitative analyses to highlight areas
   for improvement, investigating four research questions:

   o **RQ5**: On which MPAC-T items are teachers’ ratings least accurate?

   o **RQ6**: On which MPAC-T videos are teachers’ ratings least accurate?

   o **RQ7**: What difficulties did teachers encounter when rating videos?

   o **RQ8**: What difficulties did teachers encounter when using the training program?
3. **Aim 3: Describe Factors that Could Affect the Accuracy of Teachers’ MPAC-T Ratings**

This aim explores for whom and in what situations MPAC-T ratings are likely to be most accurate. Findings will inform future classroom implementation of the MPAC-T. I will use quantitative and qualitative analyses to explore two pertinent questions:

- **RQ9:** What teacher characteristics are associated with the most accurate ratings?
- **RQ10:** What factors might affect teachers’ MPAC-T ratings in the classroom?
2. METHOD

A. Computerized MPAC-T Rater Training Program

The computerized MPAC-T rater training program follows a frame-of-reference training model, which involves familiarizing raters with dimensions and standards of performance, providing examples of these dimensions and standards, and then facilitating rehearsal and feedback using this scheme (Woehr & Huffcutt, 1994). An outline of the computerized MPAC-T rater training program is included in Appendix B, and screenshots from the program are presented in Appendix C. Within each section of the MPAC-T (see Appendix A), the training program first provides a verbal explanation of each item, indicating the kinds of behaviors that would and would not merit a “check” (see Appendix C Image 2). It then presents a video example of behaviors that would merit a “check.” Next the training program asks teachers to rate example videos (see Appendix C Image 3); it then provides feedback about items that teachers rated inaccurately (see Appendix C Image 4), with optional links to repeat the training modules for those items (see Appendix B). This sequence is repeated for each section of the MPAC-T, with a final practice-and-feedback session encompassing all 20 MPAC-T items.

The computerized MPAC-T rater training program requires about 45 minutes to complete. It is a condensed version of the original, in-person MPAC rater training, which was conducted over several weeks including four, hour-long discussion sessions and up to eight hours of independent practice coding (Denham & Burton, 1996). Whereas the original MPAC training required participants to watch hours of footage at home and later discuss their ratings, the computerized program uses brief, targeted practice clips with immediate written feedback. Additionally, the original MPAC training was typically administered to new research team members (i.e., undergraduate and graduate students) who did not necessarily have experience
with early-childhood education. Therefore the original program included extensive discussions about the nature of young children’s behavior—for example, what constitutes productive involvement in early childhood. The computerized MPAC-T rater training program excludes this, as early-childhood educators are already familiar with students’ behavior.

This decrease in training time does not necessarily threaten rater performance. Observers with minimal training can reliably use some rating systems (Brotman, Gouley, & Chesir-Teran, 2005). More intensive training does not always enhance rater reliability or accuracy (Angkaw et al., 2006; LeBel et al., 2010; Lorber, 2006). Shorter, more focused rater training programs are both useful and feasible (Lorber, 2006; Tractenberg, Schafer, & Morris, 2001).

B. Participants

I recruited 22 early-childhood educators to participate in the formative evaluation of the MPACT-T rater training program. The results of a power analysis for a dependent-samples t-test suggested that 11 participants were necessary to achieve a power of .80, assuming an alpha of .05 and an effect size of .83—as reported in Woehr and Huffcutt’s (1994) meta-analysis on the effects of FOR training on rating accuracy. Assuming a more modest effect size of .60, a sample of 19 participants would be necessary. Many studies have employed a similar or smaller sample, including evaluations of rater accuracy (Engelhard, 1996; Wind & Engelhard, 2013), rater training studies (e.g., Elder, Barkhuizen, Knoch, & Von Randow, 2007; Kramer et al., 2011; Rosen et al., 2008) and even a study of computerized rater training for teachers (Hamilton et al., 2001).

1. Recruitment

Recruitment occurred in the fall of 2014. I contacted the directors and/or faculty of 79 Chicago-area early-childhood education centers and training programs. I posted recruitment
flyers in the education building at the University of Illinois at Chicago, as well as in the break room at the University of Illinois at Chicago Children’s Center. Additionally, faculty at the University of Illinois at Chicago Early Education Certification Program sent recruitment e-mails via their alumni and student listservs.

2. Screening

All interested individuals completed an online screening questionnaire, which was used to determine eligibility. To be eligible participants must have worked as an early-childhood teacher or student teacher for at least one year and must be fluent in English. Participants with hearing or visual disabilities that might prevent them from participating in the computerized MPAC-T rater training program were excluded. I also excluded participants who had taught exclusively in self-contained special education classrooms, as behaviors in these classrooms may differ substantially from those in general education classrooms.

3. Participant Characteristics

Twenty-two early-childhood educators participated. See Table 2 for rater characteristics. Most were female (96.45%) and Caucasian (68.2%). Their early-childhood education experience ranged from 1 to 22 years, averaging about six years. Most were currently employed as an early-childhood lead teacher (72.7%), and most worked in an urban school (77.3%). All held a bachelor’s or master’s degree, and most held a formal degree or certification in education (68.2%). Most teachers reported that they used assessment tools at least sometimes (90.9%). Slightly more participants taught at private preschools (54.5%) than at public ones.
C. Measure

The MPAC-T is an observational measure of preschoolers’ social-emotional competencies, operationalized as emotion expression, classroom involvement, reactions to frustration, and social interaction. To use the MPAC-T, a teacher observes one child for five minutes during free play, rating behavior as it is observed. The MPAC-T is composed of 20
### TABLE II
PARTICIPANT CHARACTERISTICS

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Percent of Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>95.45</td>
</tr>
<tr>
<td>Race</td>
<td></td>
</tr>
<tr>
<td>African-American</td>
<td>13.6</td>
</tr>
<tr>
<td>Asian</td>
<td>13.6</td>
</tr>
<tr>
<td>Caucasian</td>
<td>68.2</td>
</tr>
<tr>
<td>Other</td>
<td>4.5</td>
</tr>
<tr>
<td>Occupation</td>
<td></td>
</tr>
<tr>
<td>ECE lead teacher</td>
<td>72.7</td>
</tr>
<tr>
<td>ECE assistant teacher</td>
<td>9.1</td>
</tr>
<tr>
<td>Other</td>
<td>18.2</td>
</tr>
<tr>
<td>Highest education</td>
<td></td>
</tr>
<tr>
<td>Bachelor’s Degree</td>
<td>54.5</td>
</tr>
<tr>
<td>Master’s Degree</td>
<td>45.5</td>
</tr>
<tr>
<td>Education-Specific Training</td>
<td></td>
</tr>
<tr>
<td>Teaching Certificate</td>
<td>54.5</td>
</tr>
<tr>
<td>Bachelor’s of Education</td>
<td>18.2</td>
</tr>
<tr>
<td>Master’s of Education</td>
<td>31.8</td>
</tr>
<tr>
<td>None of These</td>
<td>31.8</td>
</tr>
<tr>
<td>Years working in ECE(^a)</td>
<td>6.25 (5.13)</td>
</tr>
<tr>
<td>School Characteristics</td>
<td></td>
</tr>
<tr>
<td>Urban</td>
<td>77.3</td>
</tr>
<tr>
<td>Suburban or Rural</td>
<td>22.7</td>
</tr>
<tr>
<td>Public</td>
<td>45.5</td>
</tr>
<tr>
<td>Private School</td>
<td>54.5</td>
</tr>
</tbody>
</table>

\(^a\)I report mean and standard deviation, in parentheses, for years working in early-childhood education. ECE = early-childhood education.
dichotomous items, each of which describes a specific behavior (see Appendix A). The teacher assigns a “check” (i.e., a rating of 1) to any item that describes a behavior the child exhibited at least once during the observation period. The teacher does not check (i.e., implicitly assigns a rating of 0 to) any items that describe behaviors the child did NOT exhibit during the observation period. Items are organized into four sections for rater convenience. Section titles are posed as questions: (1) What kinds of emotions did you see the child express?, (2) How involved was the child in class activities?, (3) When frustrated, how did the child react?, and (4) How did the child interact with others? The teacher repeats this process for each child four times over four weeks.

D. Videos

I collected videos of preschoolers’ natural behaviors during free play in their preschool classroom and outdoor play yard. All videos were collected in one classroom at a university-affiliated early-childhood education center. Videos were collected in the morning and afternoon over about one month in Spring 2014. Although multiple children are visible in each video, teachers in this study were instructed to observe and rate one focal child. The focal child was indicated by a brief, on-screen text description of his or her clothing (e.g., “yellow shirt”). Descriptive statistics pertaining to the focal children in these videos are presented in Table 3. Overall the focal children represented a mix of males and females and of Caucasian, Asian, and other ethnicities. Most videos were shot indoors. Videos depict children in a variety of settings, including a sandbox, climbing structure, block area, meal table, craft table, and writing table.
1. Training Videos

Training and assessment videos were drawn from the same pool of footage, described above. However, training videos were unique from the post-test and follow-up videos. Training videos were selected based on the experts’ ratings, described below. The MPAC-T rater training program includes two types of training videos: examples and practice.
TABLE III
ASSESSMENT VIDEO CHARACTERISTICS AND PARTICIPANT ACCURACY ON THE MINNESOTA PRESCHOOL AFFECT CHECKLIST-TEACHER

<table>
<thead>
<tr>
<th>Phase</th>
<th>Video N</th>
<th>Male</th>
<th>White</th>
<th>Asian</th>
<th>Other</th>
<th>Filmed Outside</th>
<th>Ppt. N</th>
<th>Mean (SD)</th>
<th>Min</th>
<th>Max</th>
<th>≥80% Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest</td>
<td>3</td>
<td>100</td>
<td>0</td>
<td>50.00</td>
<td>50.00</td>
<td>33.33</td>
<td>22</td>
<td>70.95</td>
<td>55</td>
<td>82</td>
<td>13.60</td>
</tr>
<tr>
<td>Training</td>
<td>24</td>
<td>54.17</td>
<td>44.44</td>
<td>44.44</td>
<td>22.22</td>
<td>25.00</td>
<td>22</td>
<td>78.05</td>
<td>68</td>
<td>86</td>
<td>36.40</td>
</tr>
<tr>
<td>Post-Test</td>
<td>10</td>
<td>40.00</td>
<td>33.33</td>
<td>50.00</td>
<td>16.67</td>
<td>20.00</td>
<td>22</td>
<td>78.05</td>
<td>68</td>
<td>86</td>
<td>36.40</td>
</tr>
<tr>
<td>Follow-Up</td>
<td>2</td>
<td>50.00</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>69.86</td>
<td>40</td>
<td>90</td>
<td>27.80</td>
</tr>
</tbody>
</table>

Note. All values are percentages. Demographics for video characteristics describe the focal child in each video. Accuracy refers to agreement with expert ratings. Pre-test and follow-up video ratings were excluded in instances where participants watched less than half of a given video. Ppt. = participant. Acc. = accuracy.
videos. Example videos are shown to explain particular behaviors, whereas practice videos are used to solicit participants’ ratings and provide feedback on those ratings. As examples I chose clips on which the raters agreed unequivocally—clips which displayed clear instances of a particular behavior. For the in-training practice opportunities I selected either clips which displayed multiple behaviors—so that raters could practice using multiple items—or rare behaviors on which participants would likely need extra practice.

2. Assessment Videos

Participants rated thirteen unique, five-minute assessment videos. They rated three videos during the pre-test, ten videos during the in-laboratory post-test, and two videos repeated from the post-test at follow-up. To avoid practice effects, none of the pre-test videos were repeated at post-test or follow-up.

E. Expert Ratings

To establish an accuracy standard I obtained MPAC-T ratings from four experts who had previously been trained to proficiency on use of the MPAC. The experts were females ranging in age from their late twenties through mid-sixties. Two developed the MPAC-R/S; the remaining two developed the MPAC-T. All experts rated all examples, practice videos, and assessment videos (pre-test, post-test, and follow-up). Experts discussed discrepancies until they reached a consensus for each item for each video. It is common to use expert consensus to develop a standard for rater accuracy (Lundh, Kowalski, Sandberg, & Landen, 2012; Muller & Dragicevic, 2003; Schleicher & Day, 1998).

F. Procedures

The study included three phases: an at-home pre-test, a laboratory session at the University of Illinois at Chicago, and an at-home one-week follow-up. Participants received $40
immediately after participating in the laboratory session and $20 after completing the follow-up, for $60 total. Six laboratory sessions were conducted; enrollment ranged from two to six participants per session.

1. **Pre-Test**

One week prior to their laboratory session participants received an e-mail containing Web links to the informed consent document and pre-test. The e-mail instructed participants to complete the pre-test before their laboratory session. Those who had not completed the pre-test two days before their laboratory session received an e-mail reminder. The pre-test first instructed participants to spend 10 minutes reading the full MPAC-T rating instrument. Next the website instructed participants to use the MPAC-T to rate three, five-minute videos. Each video was displayed on one side of the computer screen with the MPAC-T items displayed simultaneously on the other side (see Appendix C Image 1).

2. **Laboratory Session**

Each laboratory session included four components: a brief questionnaire, the online MPAC-T rater training program, a post-test, and a focus group. Participants completed the first three components—the questionnaire, training, and post-test—individually on a computer. All participants in the laboratory session then participated in the focus group together.

A. **Questionnaire.**

At the start of the laboratory session participants completed a brief computerized questionnaire to document their teaching experience, education (both general and specific to early-childhood education), and frequency of classroom assessment use.
B. **Training.**

Participants then completed the computerized MPAC-T rater training program independently while wearing headphones. They were instructed to work at their own pace. Most finished in about 45 minutes; all finished in under 60 minutes.

C. **Post-test.**

After a ten-minute break, all participants rated ten, five-minute videos using the MPAC-T. Participants used the MPAC-T on their individual computers while videos were projected on a large screen at the front of the room.

D. **Focus Groups.**

A short break followed the post-test, after which I moderated a focus group with all participants who attended a given laboratory session. An undergraduate research assistant (RA) served as assistant moderator. The focus group guide is included in Appendix D. Focus groups were video recorded and participants were identified by unique numbers placed in front of them. Focus groups followed a semi-structured format, with main questions and probes to evoke greater detail or follow up on relevant information. Topics included the MPAC-T rater training program, the experience of rating videos using the MPAC-T, and potential classroom use of the MPAC-T.

3. **Follow-Up.**

At the end of the laboratory session participants were told to expect an e-mail one week later. This e-mail contained a link to the online follow-up assessment, which instructed participants to rate two, five-minute videos using the MPAC-T (see Appendix C Image 5). Participants who did not complete the follow-up assessment within one week received a reminder e-mail. The follow-up videos were chosen from among the ten post-test videos to
represent a wide array of MPAC-T behaviors. Other studies have implemented a similar method, assessing test-retest reliability by asking participants to rate the same videos several weeks apart (Rodger et al., 2014; Allen et al., 2012).

G. **Mixed-Methods Research Design**

This study employs a convergent mixed-methods research design, in which quantitative and qualitative data are collected during the same phase of the research project and merged to jointly address study aims (Creswell, Klassen, Clark, & Smith, 2011). Following the recommendations of Creswell and colleagues (2011), study aims were driven by substantive questions rather than research methods. Quantitative and qualitative findings were merged to address Aims 2 and 3. For these aims qualitative codes were developed with the explicit goal of elaborating on quantitative findings. However, I report all themes that emerged under these codes, whether they supported or contradicted the quantitative findings. Other studies have likewise used both quantitative analyses to describe raters’ ratings and qualitative analyses to describe factors that influence their ratings (e.g., Kim, 2009; Winke, Gass, & Myford, 2011). In this way the present study addresses a call for qualitative analyses of raters’ thought processes (Johnson & Lim, 2009).

H. **Focus Group Transcription, Coding, and Reliability**

Undergraduate RAs transcribed focus groups from videos, attributing each participant’s comments to his or her unique number, which was displayed in the video. Transcripts were stripped of all references to individuals not participating in the study.

Because I already had ideas about the aspects of the rating context which might affect teachers’ ratings—specifically training, the rating process itself, and the classroom context—I used a directed approach to expand upon these codes: a method outlined by Hsieh and Shannon
I used the a-priori research aims (i.e., Aims 2 and 3) to guide the generation of major codes, but I allowed themes within those codes to emerge from the content of the focus groups.

Coding took place after all focus groups were completed. Coders were myself, a female in her twenties, and one RA, a male in his twenties. We coded transcripts using Nvivo qualitative data analysis software (Version 10; QSR International, 2012). To provide context for each comment, the unit of analysis was the utterance—that is, one speaker’s statement from the time he or she began talking until someone else began speaking. The codebook began with major codes derived from my research aims. Revisions to the codebook—primarily to definitions of these major codes—followed an iterative process. After reading the initial codebook the RA and I coded one transcript independently, discussed discrepancies, and revised code definitions accordingly. We repeated this process until our mean inter-rater Kappa surpassed .70. The final coding structure is described in Table 4. After we revised the codebook and revised our codes based on our discussions, Kappa for the major codes was .75, which is considered substantial agreement (Landis & Koch, 1977). After this I coded the remaining three transcripts independently.

Percent coverage for each code and theme was calculated as the percent of characters (including spaces) across all six focus groups which were coded with a given code or theme. Coverage values throughout are low, as the focus group guide included additional questions for potential use in future studies (e.g., “Imagine you needed to complete the training program on your own, at home or at school. How would your training experience be different?”)
### Table IV
#### Teacher Focus Group Codes and Themes

<table>
<thead>
<tr>
<th>Code</th>
<th>Theme</th>
<th>Kappa</th>
<th>Percent of Participants</th>
<th>Percent Coverage</th>
<th>Focus Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Difficulties When Rating Videos</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-test length</td>
<td></td>
<td>0.78</td>
<td>95.45</td>
<td>33.84</td>
<td>6</td>
</tr>
<tr>
<td>Measure specificity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Specific Items</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.-1b. Positive emotions*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Negative emotions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Engaged and engaged</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Emotionally absent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Fidgety and uninterested*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Talking about frustration</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9a. Frustration: Neutral/Positive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Displayed leadership skills</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13 &amp; 15. Took turns and shared*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Cooperated with peers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16. No interactions for 3 minutes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Difficulties During the Training Program</strong></td>
<td></td>
<td>0.67</td>
<td>90.91</td>
<td>13.69</td>
<td>6</td>
</tr>
<tr>
<td>Need more detail</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technical glitches</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Rating in the Classroom</strong></td>
<td></td>
<td>0.80</td>
<td>72.72</td>
<td>10.19</td>
<td>6</td>
</tr>
<tr>
<td>Practical barriers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge of the child</td>
<td></td>
<td>27.27</td>
<td>3.29</td>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>

*Note.* If a particular item is not listed it was not mentioned as difficult to use. Percent coverage indicates the percent of all characters (including spaces) attributed to a particular code or theme. The focus groups column indicates the number of focus groups in which at least one participant mentioned a given code or theme. * = Quantitative analyses indicated that teachers’ ratings of this item were particularly inaccurate.
3. RESULTS

A. Completion Rates

All participants rated all post-test videos in the laboratory. However, some participants failed to rate some videos during the at-home pretest and follow-up phases. Of 330 total video ratings, 10 were missing due to non-compliance.

I omitted 12 additional video ratings due to short viewing time. The pre-test and follow-up interface measured how long participants spent on the webpage for each video. I omitted a participant’s ratings for a given video if he or she spent less than half of the video length—specifically, 2 minutes and 30 seconds—on the video’s page. This procedure was not necessary for post-test videos, as they were screened at full length in the laboratory.

B. Aim 1: Assess the Efficacy of the MPAC-T Rater Training Program

1. RQ1: How does Rater Training Influence the Accuracy of Teachers’ MPAC-T Ratings?

A teacher’s rating—specifically, either a check or no check for a single MPAC-T item—was considered accurate if it matched the experts’ rating. To explore whether the accuracy of teachers’ ratings improved with training, I conducted a dependent-samples t-test, comparing pre- and post-test accuracy. Overall, the teachers’ rating accuracy was significantly higher after training ($M = 78.05, SD = 5.63$) than before training ($M = 70.95, SD = 6.22$), $t (21) = 5.07, p < .001$.

2. RQ2: How Accurate are Teachers’ MPAC-T Ratings after Training?

At post-test the teachers’ ratings agreed with experts’ ratings on 78% of MPAC-T items (SD = 5.29). As in other rater training studies, I used 80% accuracy as a criterion for overall rater
accuracy (Kobak et al., 2007; Kobak, Stone, Ousley, & Swanson, 2011). About one-third of teachers achieved or surpassed this 80% accuracy criterion at post-test (see Table 3).

I also computed Cohen’s Kappa, a chance-corrected measure of inter-rater agreement, to compare each teacher’s ratings to the experts’ ratings. Kappa values can range from -1 to 1. A Kappa value of 1 indicates complete agreement, 0 indicates chance agreement, and -1 indicates complete (systematic) disagreement. Teachers’ Kappa values ranged from .13 to .60; see the third column of Table 5 for the Kappa distribution. Most teachers assigned ratings that agreed moderately with the experts’ ratings; that is, 64% of teachers had Kappa values between 0.41 and 0.60 (Landis & Koch, 1977).

3. RQ3: How Internally Reliable are Teachers’ MPAC-T Ratings after Training?

It is important for teachers to exhibit intra-rater reliability on the MPAC-T, because teachers may wish to compare their ratings across items, children, or time points. Intra-rater reliability ensures that these comparisons will reflect true differences, rather than rater error.

To determine teachers’ post-test intra-rater reliability I used the Facets (v3.67.0) software to run multi-facet Rasch measurement (MRFM) analyses on teachers’ raw ratings (Linacre, 2010). Since MFRM assumes a unidimensional measure, I reverse-coded the negatively oriented MPAC-T items (see Table 1). I examined three facets of the data: child social-emotional competence (by video), teacher severity (i.e., a teacher’s tendency to assign low or high ratings), and item difficulty (i.e., how easy or hard it is to receive a score of 1 on a particular item). For this multi-facet Rasch analysis I used a partial credit model, in which each rating scale for each item was modeled separately. See Appendix E for the MRFM model, and Table 6 for summary statistics.
As indices of intra-rater reliability I used the teacher infit and outfit mean-square statistics from the output of my MRFM analysis. Both mean-square fit statistics indicate the degree to which each teacher contributes meaningful information to the measurement of social-
## TABLE V
### DISTRIBUTION OF TEACHER KAPPA VALUES

<table>
<thead>
<tr>
<th>Kappa Range</th>
<th>Agreement</th>
<th>Accuracy</th>
<th>Test-Retest Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01-0.20</td>
<td>Slight</td>
<td>4.55%</td>
<td>0</td>
</tr>
<tr>
<td>0.21-0.40</td>
<td>Fair</td>
<td>31.82%</td>
<td>33.33%</td>
</tr>
<tr>
<td>0.41-0.60</td>
<td>Moderate</td>
<td>63.64%</td>
<td>28.57%</td>
</tr>
<tr>
<td>0.61-0.80</td>
<td>Substantial</td>
<td>0</td>
<td>33.33%</td>
</tr>
<tr>
<td>0.81-0.99</td>
<td>Almost Perfect</td>
<td>0</td>
<td>4.76%</td>
</tr>
</tbody>
</table>

*Note.* Accuracy and test-retest reliability values represent the percent of teachers whose Kappa values fell within each range. *a* Landis & Koch, 1977. *b* Agreement between teachers’ ratings and experts’ ratings; N = 22. *c* Agreement between an individual teacher’s post-test and follow-up ratings for the same video; N = 21.
TABLE VI
SUMMARY STATISTICS FOR MULTI-FACET RASCH MEASUREMENT ANALYSES OF MPAC-T INTRA-RATER RELIABILITY AND ACCURACY

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Analysis of Intra-Rater Reliability&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Analysis of Accuracy&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Teaching Experience</th>
<th>Degree</th>
<th>Assessment Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>1.23</td>
<td>1.60</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SD</td>
<td>0.39</td>
<td>1.06</td>
<td>0.34</td>
<td>0.08</td>
<td>0.12</td>
</tr>
<tr>
<td>N</td>
<td>22</td>
<td>20</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Infit mean-square</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.01</td>
<td>0.97</td>
</tr>
<tr>
<td>SD</td>
<td>0.08</td>
<td>0.05</td>
<td>0.08</td>
<td>0.02</td>
<td>0.04</td>
</tr>
<tr>
<td>Min</td>
<td>0.88</td>
<td>0.93</td>
<td>0.88</td>
<td>0.99</td>
<td>0.94</td>
</tr>
<tr>
<td>Max</td>
<td>1.17</td>
<td>1.11</td>
<td>1.14</td>
<td>1.02</td>
<td>1.02</td>
</tr>
<tr>
<td>Separation statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Separation Reliability</td>
<td>0.79</td>
<td>0.93</td>
<td>0.87</td>
<td>0.05</td>
<td>0.81</td>
</tr>
<tr>
<td>(\chi^2) (df)</td>
<td>101.9* (21)</td>
<td>169.2* (19)</td>
<td>69.7* (9)</td>
<td>2.3 (2)</td>
<td>5.3* (1)</td>
</tr>
</tbody>
</table>

Note. <sup>a</sup> Multi-facet Rasch measurement analysis using raw ratings. <sup>b</sup> Multi-facet Rasch measurement analysis using scored ratings (1 = agreement with experts, 0 = disagreement with experts). MPAC-T = Minnesota Preschool Affect Checklist-Teacher. \(\chi^2\) = fixed (all-same) chi-square statistic. * p < .05.
emotional competencies. Infit mean-square is an inlier-sensitive statistic that describes teacher’s overall performance. By contrast, outfit mean-square is an outlier-sensitive statistic that is sensitive to rare, highly unexpected ratings. Interpreting both statistics in tandem can help to distinguish whether a rater’s unexpected ratings were quite prevalent (indicated by a high infit mean-square statistic) or more rare (indicated by an acceptable infit mean-fit square statistic and a high outfit mean-square statistic).

The expected value for a mean-square fit statistic is 1. Mean-square fit statistics range from 0 to infinity. Because the MPAC-T is a low-stakes formative assessment tool, an acceptable range for the mean-square fit statistics is 0.5 to 1.5 (Wright & Linacre, 1994); ratings below or above this range do not provide productive information for measuring social-emotional competencies. A teacher infit or outfit mean-square statistic below 0.5 would indicate that the teacher may not have rated each item independently, but instead may have tended to assign many of the same ratings across items. This could indicate that the teacher was unable to distinguish among items. By contrast, a teacher infit or outfit mean-square statistic above 1.5 would indicate that at least one of the teacher’s ratings was quite inconsistent with the overall pattern of ratings (Linacre, 2002).

For this Rasch analysis of teacher invariance, teacher infit mean-square statistics ranged from 0.88 to 1.17, within the acceptable range (see Table 6). This suggests that overall the teachers were able to use the MPACT in an internally consistent fashion. However, teacher outfit-mean-square statistics ranged from .58 to 2.3; four teachers had outfit-mean-square statistics above the upper limit of 1.5. While the overall pattern of these teachers’ ratings was largely consistent, indicated by acceptable infit mean-square statistics, high outfit mean-square
statistics suggest that these four teachers occasionally assigned ratings that were quite unexpected.

I used the table of unexpected responses from the output of my MFRM analysis to identify patterns in these four teachers’ unexpected ratings. Among these four teachers, Video 10, Yellow Shirt, received the most unexpected ratings. This video was shown last during the post-test; perhaps these four teachers were particularly susceptible to fatigue. Alternately, there may be something unique about Video 10 that is inconsistent with the overall pattern of social-emotional competencies depicted in the rest of the videos. Among these four teachers, Item 7 (“Reacted to a frustrating situation with interpersonal aggression”) received more unexpected ratings than any other item. Interestingly, most of these four teachers’ unexpected ratings for Item 7 were assigned on Video 10—Yellow Shirt. Specifically, for these four teachers did not “check” Item 7 for the Yellow Shirt video, contrary to the Rasch model’s prediction. Video 10 may depict, or not depict, frustrated interpersonal aggression in a way that is not adequately addressed by the MPAC-T rater training program.

4. RQ4: How Stable are Teachers’ MPAC-T Ratings One Week after Training?

Test-retest reliability is one indicator that a teacher’s ratings of the same child over the school year would reflect true development, rather than rater error. To assess test-retest reliability I compared teachers’ ratings of the same two videos at post-test and at one-week follow-up. One teacher is excluded from these analyses because she did not submit any follow-up ratings (n = 21).

First I calculated item-level percent consistency, or the number of teachers whose rating for each item remained the same from post-training to follow-up. For the first follow-up video, on average teachers’ ratings were 80.63% consistent (SD = 16.64, n = 16) from post-training to
follow-up. For the second follow-up video, on average, teachers’ ratings were 76.41% consistent ($SD = 13.04$, $n = 12$). I also computed Cohen’s Kappa to compare each teacher’s ratings for these two videos from post-test to follow-up as a chance-corrected measure of test-retest reliability. Test-retest Kappas ranged from .21 to .83; see the fourth column of Table 5 for the full distribution. Kappa values were distributed about evenly among fair, moderate, and substantial agreement.

5. **Aim 1 Summary**

Teachers’ accuracy improved significantly from pre-test to post-test. After completing the MPAC-T rater training program teachers demonstrated 78% rating accuracy as compared with experts. Kappa values indicated moderate agreement with experts. Overall, teachers’ ratings were internally consistent. However, unexpected responses for four teachers suggest that Video 10 and Item 7 may require closer consideration. Teachers’ ratings exhibited fair to substantial test-retest reliability.

C. **Aim 2: Identify Areas for Improvement in the MPAC-T Rater Training Program**

**Quantitative analyses.** To provide focused diagnostic information about the accuracy of ratings for teachers, MPAC-T items, and videos (i.e., to answer RQs 5, 6, and 9), I conducted a second MFRM analysis. In the first MFRM analysis, used to answer RQ3, I used raw ratings to examine teacher invariance; in this second analysis I used scored ratings (i.e., 1 if the teacher’s rating agreed with the experts’ rating, 0 if the teacher’s rating disagreed with the experts’ rating) to investigate the accuracy of teachers’ ratings, as compared with the experts’ ratings (Engelhard, 1996; Engelhard, Davis, & Hansche, 1999; Wind & Engelhard, 2013). This allowed me to directly compare how accurate ratings were for individual teacher characteristics, items, and videos. I used a partial credit model, which modeled each rating scale for each item separately.
To better understand the characteristics driving differences in teacher accuracy, I replaced the teacher facet with three teacher demographic characteristic facets: years of teaching experience (i.e., 0-5, 6-10, or over 10), frequency of classroom assessment use (i.e., rarely, sometimes, often, or all the time), and highest level of education (i.e., bachelor’s or master’s degree). I used the results from this analysis to investigate relationships between these teacher characteristics and rating accuracy. See Appendix E for the MRFM model. Infit mean-square values for all facets fell within the acceptable range of 0.5 to 1.5, indicating sufficient fit to the Rasch model (see Table 6). This finding suggested that it was appropriate to move forward to draw meaningful conclusions from the Rasch-generated accuracy measures.

Multi-facet Rasch measurement analysis output includes measures (with standard error estimates) of the accuracy of ratings assigned by each teacher characteristic, for each video, and for each item. The three teacher characteristic facets and the video facet were centered at zero (i.e., the item facet was the only non-centered facet). All facets were positively oriented. In other words, teachers with higher accuracy measures tended to assign more ratings that agreed with the experts’ ratings than teachers with lower accuracy measures. Similarly, items (and videos) with higher accuracy measures were ones that showed higher agreement in the ratings of the experts and the teachers. Table 6 displays summary statistics for MPAC-T accuracy scores.

1. **RQ5: On Which MPAC-T Items are Teachers’ Ratings Least Accurate?**

Low item accuracy may indicate a need for item or training revision. Item accuracy measures ranged from 0.55 to 5.51 logits. The results from the fixed (all-same) chi-square analysis indicated that at least two of the items differed in terms of how easy it was for teachers to assign accurate ratings, $\chi^2 (19) = 172.3$, $p < .01$. The high item separation reliability (0.93) indicated that teacher accuracy varied across the 20 items. The strata estimate was 5.23,
suggesting that among the 20 items there were about five statistically distinct levels, or groupings, of items that differed in terms of how easy it was for teachers to assign accurate ratings. These five item-accuracy groupings are evident in the item map (Figure 1); the more accurately an item was rated, the closer it is to the top of the map. The six items in the grouping closes to the bottom of the map (1, 1a, 4, 6, 13, and 15) warrant further inspection since these were the items that were most difficult to rate accurately.

To gain a better understanding of what made these six low-accuracy items hard to rate, I computed simple counts of under-rating—a teacher scored an item as 0 (not checked) when the experts scored it as 1 (checked), and over-rating—a teacher scored an item as 1 (checked) when the experts scored it is as 0 (not checked). See Figure 2. Teachers tended to over-rate these low-accuracy items (i.e., they checked them when the experts did not). One exception was Item 4 (“Was involved in an activity that he/she helped to organize”), which was under-rated; teachers failed to check this item as frequently as experts did.

2. **RQ6: On which MPAC-T videos are teachers’ ratings least accurate?**

A video accuracy measure indicates the likelihood of the teachers assigning ratings that matched those that the experts assigned to that particular video; these measures ranged from -0.01 to 1.49. Videos with higher measures were easier for teachers to rate accurately. The fixed (all-same) chi-square test indicated that at least two of the videos differed in terms of how easy it was for teachers to rate them accurately, \( \chi^2 (9) = 71.6, p < .01 \). The video separation reliability was .87, indicating that teacher accuracy varied across the ten videos. The strata estimate was 3.84, suggesting that among the ten videos there were about four statistically distinct levels, or groupings, of videos that differed in terms of how easy it was for teachers to assign accurate ratings. Ratings were least accurate for the three final videos screened in the laboratory session.
Accuracy scores decreased with video order from the eighth to the tenth video. This may suggest a teacher fatigue effect.
Figure 1.
Variable map showing the ordering of the accuracy measures for MPAC-T teacher characteristics (teaching experience, education, assessment use), videos, and items. Higher accuracy measures represent more ratings that agreed with those assigned by experts.

<table>
<thead>
<tr>
<th>Accuracy Measure</th>
<th>Teaching Experience</th>
<th>Education</th>
<th>Assessment Use</th>
<th>Video</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>More Accurate Ratings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Fewer Accurate Ratings | | | | | |
|-------------------------|-----------|----------------|-----|------|
| Rarely | 1 | 4 | 5 |
| 6-10yrs | Bachelor’s | | | | |
| >10yrs | 1-5yrs | Master’s | | | |
| | Sometimes | All The Time | Often | | |
| | 6 | 2 | 3 |
| | 8 | | 9 | 10 |

Figure 2.
Rates of under- and over-rating for MPAC-T items and videos.
To gain a better understanding of what made these three low-accuracy videos hard to rate, I computed simple counts of under-rating and over-rating. See Figure 2. The results mirrored those for the low-accuracy MPAC-T items; teachers tended to over-rate these low-accuracy videos. That is, on these three final post-test videos teachers “checked” more items than the experts did.

3. RQ7: What Difficulties Did Teachers Encounter When Rating Videos?

I used a qualitative analysis of teachers’ focus group responses to answer this question; see Table 4 for the coding structure. This code was defined by two criteria:

1. The comment must pertain to the MPAC-T measure itself or to the post-test phase (i.e., coding 10 videos) of the study.
2. The comment must pertain either to criticisms (e.g., “I don’t like…”) or potential future improvements (e.g., “It would be helpful if…”).

Teachers in all six focus groups (33.85% coverage) mentioned difficulties that they experienced when rating videos at post-test. Their comments ranged from thoughts about the rating process as a whole to specific concerns about individual items.

A. Post-test length.

Five teachers (1.99% coverage) mentioned that the post-test phase—which consisted of 10, 5-minute videos—was too long. This comment exemplifies teachers’ general sentiment: “I kind of glazed over part of the time because it’s 10 videos so it’s a little long” (1633). This may explain the quantitative finding that teachers’ accuracy declined over the final three post-test videos (see RQ6); teachers may have begun to lose interest.

B. Measure specificity.
Six teachers across five focus groups (5.74% coverage) mentioned that the MPAC-T measure was not specific or detailed enough for them to use it accurately.

1622: From reading it, it’s not as detailed as when it was being explained during the training program. So at the end when I was trying to use it with the video, there was some parts where I was trying to remember what it was, how it was more detailed during the training part.

This comment highlights a sentiment echoed by other teachers: they wished more details from the training program were included on the measure itself. To better understand the kinds of details teachers would have liked the MPAC-T to provide, I then explored teachers’ comments about specific difficult items.

C. Specific items.

Teachers described difficulties using particular items to rate the post-test videos. I will first present themes pertaining to the six items distinguished as low-accuracy in the quantitative results (RQ5): Items 1, 1a, 13, 15, 4, and 6. I will then present findings about additional items.

Positive emotions. Eleven teachers (50% of the sample) described difficulties rating the items pertaining to positive emotions (i.e., 1 through 1b). They often mentioned the struggle to distinguish positive from neutral affect:

1611: The ‘showing positive emotions in any manner,’ so I know they gave an example of that or a definition of that, but if they have no affect... do we say that that's kind of positive or is it negative? That was hard to decipher.

This statement reflects an inclination to infer positive affect when it is not directly observable.

Others echoed this sentiment, suggesting that cultural factors might create discrepancies between a child’s true feelings and his or her observable affect:

1634: I found it easier to code behaviors than to guess about the kids’ emotions. Especially, one of the things I was thinking, since there were so many Chinese children at that school, that since there’s cultural differences in how kids express emotions you
know, I wasn’t completely convinced that they weren’t showing positive emotion or negative emotion.
Thus teachers may have over-rated Items 1 and 1a—as indicated in the quantitative results (RQ5)—because they were inferring positive emotions that were not directly observable.

Sharing and turn-taking. Fifteen teachers spanning every focus group (5.80% coverage) mentioned difficulties distinguishing between items 13 (“Took turns with classroom materials,”) and 15 (“Shared toys or other materials”), both of which were identified as low-accuracy items in the quantitative analyses (RQ5). This teacher succinctly stated the confusion that many expressed: “Being in the classroom you think like, as long as they’re like using the same stuff, you can consider that sharing or taking turns” (1656). This conceptualization, “using the same stuff,” is broad enough to encompass both sharing and turn-taking, as well as behaviors that are neither sharing nor turn-taking, such as climbing on the same structure. This may explain why teachers over-rated both items.

Fidgety and uninterested. Six teachers in four focus groups mentioned difficulties rating Item 6 (“Looked fidgety and uninterested or uninvested in the activity but still ‘tuned in’ and aware of his/her surroundings”). Teachers stated that a child could be interested in an activity even if his or her outward appearance were fidgety or aloof. For example, a child who has difficulty remaining still might be interested in an activity nonetheless. Further, a child who is uninterested in his or her own activity might actually be quite interested in another activity elsewhere, as expressed in this comment: “What if they are switching, is it still the activity? Cause then they’re uninterested in the one that they were doing but now they’re interested in someone else’s, you know?” (1633). Teachers’ comments suggest that they felt Item 6 did not apply in many situations because children exhibited only some aspects of the item. This suggests
that teachers would select Item 6 less often than the experts did, because teachers often felt it did not apply. However, the data do not support this (see Figure 2).

*Involved in a self-organized activity.* Although the quantitative analyses (RQ5) revealed poor teacher accuracy on Item 4 (“Was involved in an activity that he/she helped to organize”), teachers did not mention this item as difficult to rate. While the other low-accuracy items were usually over-rated (see Figure 2), Item 4 was typically under-rated. These findings could suggest that teachers were unaware of Item 4; they did not think to rate it or to mention it.

*Emotionally absent.* In addition, teachers mentioned difficulties with seven items which they did not rate particularly inaccurately (see Table 4). Of these, Item 5 (“Looked ‘emotionally absent’ and showed no involvement in an activity”) was most commonly mentioned; four teachers discussed it. They expressed uncertainty about whether a child was truly emotionally absent, as exemplified in this interchange:

1612: I was kind of confused about the 'emotionally absent' part to it. I wish there was some sort of time indicator, like this kid is emotionally absent for 30 seconds. Because a lot of times when kids are intimidated they'll just kind of stare off into the distance, but does that mean they're emotionally absent or does that mean they're thinking?
1611: Or are they tired?
1613: Yeah.
1612: Tired, yeah. Or daydreaming, and that's part of the process especially if you're doing creative activities, which is what you're doing most of the day in pre-k or in early childhood classrooms.

These comments echo those regarding Item 6 (“fidgety and uninterested”): They express the belief that a child’s observable behaviors might incorrectly convey emotional absence. A few teachers also described difficulties with additional items (see Table 4). However these difficulties seem relatively idiosyncratic; each was mentioned by only three or fewer teachers, their coverage was less than 1%, and they had average quantitative accuracy measures (RQ5).
4. **RQ8: What Difficulties Did Teachers Encounter When Using the Training Program?**

Twenty teachers (13.69% coverage) mentioned aspects of the MPAC-T rater training program that could be improved. Comments under this code met two criteria:

1. They were reactions to the 45-minute computerized training program, including the explanation slides with voiceover, video examples, opportunities for practice, and written feedback.
2. They described either suggested revisions to the training program (e.g., more detail) or difficulties that a teacher experienced (e.g., technical problems).

Teachers’ comments under this code focused primarily on the need for more detail in the program and technical glitches.

**D. Need for more detail.**

Eight teachers (3.97% coverage) described a need for more detail, explanation, or examples in the training program. Many mentioned lingering questions about how behaviors were defined or why video examples were applicable, as in this quote:

> 1665: The videos were helpful but I think that the way it was set up, they would tell you what the descriptor was and then they would show you the video and they would move on. And it would have been good before they moved on to go back and be like, ‘In this video the reason why it fits into this category is because the little boy did this and this and this and this is why it fits.’ Because there were a few I was a little unsure of like why it would have fit, so it’s nice to have that immediate feedback like right after you see it to really understand it.

Other teachers described wanting more opportunities for practice and feeling generally that the training program did not include enough detail.

**E. Technical glitches.**

Eight teachers (2.31% coverage) mentioned technical difficulties, although they did not all describe the same problem. This teacher describes the most common difficulty, in which an explanation video paused or skipped: “Like halfway through it stopped and went back to the very
beginning. I was like, ‘What? That’s so weird.’ It was almost every one of them too where it went back after halfway” (1662). This teacher is referring specifically to explanatory videos about MPAC-T items, rather than to the post-test videos. At best this glitch may have annoyed teachers, who had to relocate their place in the video. At worst these glitches may have caused teachers to miss training content. It is not possible to determine from these data how many teachers experienced technical glitches, as some may have chosen not to discuss glitches during focus groups.

5. **Aim 2 Summary**

Five items—1, 1a, 4, 6, 13, and 15—were rated least accurately. Teachers tended to “check” these items more often than the experts did. Ratings were least accurate for the final three videos shown during post-test. Teachers tended to check more items for these videos than the experts did. Focus group responses indicated that when rating videos, teachers felt the post-test was too long, wished the measure were more specific, and had difficulties using particular items—especially the low-accuracy items indicated in the quantitative results. Further, teachers reported areas for improvement in the MPAC-T rater training program, including lack of detail and technical glitches.

D. **Aim 3: Describe Factors that Could Affect the Accuracy of Teachers’ MPAC-T Ratings**

1. **RQ9: What Teacher Characteristics are Associated with the Highest Accuracy?**

I used the multi-facet Rasch measurement analysis of rater accuracy described under Aim 2 to answer this question. For teaching experience the fixed (all-same) chi square was nonsignificant, $p = .32$. That is, the average accuracy measures for the three groups of teachers (i.e., those with 1-5 years of experience, 6-10 years of experience, and more than 10 years of experience) were not statistically significantly different. However, as a group, teachers with a
bachelor’s degree (average accuracy measure = 0.92, n = 12) tended to rate videos more accurately than those with a master’s degree (average accuracy measure = 0.64, n = 10 ), \( \chi^2 (1) = 5.3, p < .05 \).

Accuracy also varied significantly by teacher-reported frequency of assessment use, \( \chi^2 (3) = 12.4, p < .05 \). Accuracy measures for the four levels or frequencies of assessment use ranged from 0.45 to 1.32. I used independent-samples t-tests to determine which pairs of accuracy measures (e.g., “rarely” vs. “sometimes”) were statistically significantly different from one another. Teachers who reported that they used assessments rarely (average accuracy measure = 0.33, n = 2) had significantly higher accuracy measures than teachers who reported that they used assessments often (average accuracy measure = -0.20, n = 7), \( t(7) = 3.39, p < .05 \), or all the time (average accuracy measure = -0.09, n = 6), \( t(6) = 2.68, p < .05 \). These results indicate that teachers who rarely use assessments were able to assign more accurate ratings on the MPAC-T than those who use assessments more frequently.

2. RQ10: What Factors Might Affect Teachers’ MPAC-T Ratings in the Classroom?

This qualitative code was defined as a teacher’s explicit description of ways in which rating in the classroom would differ, either positively or negatively, from rating video in the laboratory (as in the present study). Fifteen teachers (68.18% of the sample) discussed factors that would affect their MPAC-T ratings in the classroom. The most common themes were practical barriers and prior knowledge of the focal child.

A. Practical barriers.

Ten teachers (4 focus groups, 3.79% coverage) mentioned practical barriers to using the MPAC-T in their classrooms. Most described the difficulties of observing one child for five minutes while supervising an entire classroom, as this teacher explains: “I think it would be
harder in the classroom because you do have so much other stuff going on” (1641). One teacher also mentioned the burden of conducting multiple observations per child, as the MPAC-T is intended to be used:

1665: And even on a longer term basis, like if you want to say, ‘I’m going to do this assessment on my kids multiple times within a few months so that I could get a better snapshot,’ then five minutes on like 20 kids in a week… you really can’t do that because it just does take so long.

When asked about possible solutions to these challenges, most teachers suggested that others, such as such as teachers from other classrooms, could conduct the observations.

B. Knowledge of the Child.

Six teachers (3.29% coverage) mentioned ways in which their prior knowledge of the child might influence their MPAC-T ratings. None of the teachers suggested that prior knowledge would make the rater process more difficult or less accurate. Rather, most thought prior knowledge would make the rating process easier because teachers would understand both the context of the behavior and the child’s unique expression style. This teacher describes how prior knowledge would help her to rate the first section of the MPAC-T—emotional expression.

1622: I think I might be able to read more easily at least the first portion of it. Just because knowing, happening to know some of the kids in the video now, I could hear them and like depending on the situation, I could tell, ‘Okay that might look like a negative reaction but that’s actually like a neutral or positive one for him.’ So probably classroom would be easier and quicker to rate with children that I know better.

This comment implies that the teacher might make inferences about a child’s emotions based on her prior knowledge of the child’s emotional tendencies.
3. **Aim 3 Summary**

Teachers with bachelor’s degrees assigned more accurate ratings than those with master’s degrees. Similarly, those who used assessments less frequently in their own classrooms tended to assign more accurate ratings. Accuracy did not differ by teaching experience. Teachers described how their MPAC-T ratings would be affected by practical barriers and prior knowledge of the focal child.
4. DISCUSSION

I conducted a formative evaluation of a computerized program to train early-childhood educators as raters on the Minnesota Preschool Affect Checklist—Teacher. This mixed-methods evaluation is part of an iterative process; results will be used to improve the MPAC-T rater training program, with the ultimate goal of supporting teachers’ assessment of preschoolers’ social-emotional competencies. Three major findings emerged from the study: the MPAC-T rater training program is moderately effective, it can be improved in several ways to increase user experience and effectiveness, and practical barriers must be addressed before the MPAC-T is implemented in classrooms. I will discuss each finding along with possible future directions and limitations of the present study.

A. **Efficacy of the MPAC-T Rater Training Program**

   1. **Trained Teachers Can Use the MPAC-T Accurately**

      The accuracy of teachers’ MPAC-T ratings—that is, the degree to which they agreed with experts’ ratings—improved from pre-test to post-test. Likewise, a prior study found that teachers who receive rater training tend to rate students’ behaviors more accurately, as compared with untrained controls (LeBel et al., 2010). While many early-childhood social-emotional assessments require extensive training (Epstein, Synhorst, Cress, & Allen, 2009), the present study demonstrates that even a brief rater training program can improve the accuracy of teachers’ social-emotional ratings. This is a welcome finding, as the intensive training requirement for the original MPAC was a significant impediment to its use in schools (Denham et al., 2010). The present findings provide initial evidence that MPAC-T rater training can be both condensed and effective.
Teachers’ average post-test accuracy (i.e., agreement with experts) was 78%. This falls slightly below the 80% accuracy criterion implemented in other rater training studies (Kobak et al., 2007, 2011). However, this 80% accuracy criterion has traditionally been applied to participants’ scores on multiple-choice tests about measure use (e.g., vignette responses) rather than to their actual ratings assigned using the measure upon which they were trained. The limited response options on a multiple-choice test may have inflated participants’ accuracy rates, making 80% a more realistic accuracy criterion in prior studies than in this one.

Correcting for chance, most teachers’ ratings agreed moderately with ratings assigned by experts. This is consistent with other studies reporting low to moderate agreement between teachers and independent observers (Achenbach, McConaughy, & Howell, 1987; Milfort & Greenfield, 2002; Phillips & Lonigan, 2010). Some disagreement between teachers and experts may be inevitable; informant discrepancies have been described as “one of the most robust findings in child clinical research” (De Los Reyes & Kadzin, 2005, p. 483). However, targeted revisions to the MPAC-T rater training program, described below, may be instrumental in increasing agreement between teachers’ and experts’ ratings on the MPAC-T.

2. Teachers’ MPAC-T Ratings are Reliable

Intra-rater reliability indicates that cross-child or longitudinal comparisons (e.g., comparing ratings of one child assigned throughout the schoolyear) among MPAC-T ratings reflect true differences, rather than rater error. Overall, teachers’ post-training ratings were internally reliable. Similarly, Weigle (1998) found that rater training produced more internally consistent ratings. However, a few teachers occasionally assigned highly unexpected ratings on the MPAC-T. This information is useful for diagnosing areas for improvement, but it does not suggest a significant flaw in teachers’ use of the MPAC-T. Rather, these unacceptable outfit
mean-square values often flag ratings that are “lucky guesses and careless mistakes” (Linacre, n.d.). However, the fact that multiple teachers assigned unexpected ratings for Item 7 (interpersonal aggression) in Video 10 (Yellow Shirt) suggests that there may be something unique about the depiction of interpersonal aggression in that particular video. The MPAC-T rater training program may require revisions to more directly address this particular depiction so that teachers understand how to rate similar behavior in their classrooms.

Teachers rated the same two videos at post-test and at one-week follow-up. Their ratings were fairly to substantially consistent. Likewise, prior studies reported moderate test-retest reliability of ratings on videos repeated after a delay (Allen et al., 2012; Rodger et al., 2014). Because teachers were retested on identical videos, this finding reflects the true stability of teachers’ ratings over time; it is not influenced by any changes in child behavior from pre-test to follow-up. While MPAC-T test-retest reliability was acceptable on the whole, it was stronger for some teachers than for others. Future studies could explore individual predictors of test-retest reliability. Attention or effort at follow-up may be influential. Because follow-up ratings were completed at home, some teachers may have had other stimuli competing for their cognitive resources.

B. Areas for Improvement in the MPAC-T Rater Training Program

1. Provide More Detailed Information about Items

Teachers reported that the MPAC-T rater training program lacked sufficient detail (RQ8). Future iterations of the training program could include explanations of examples after they are shown, as one teacher suggested. The training program could also provide more opportunities for practice and feedback, as others requested. Indeed, teachers respond positively to feedback on
their ratings (Elder et al., 2005) and learn more from practice with feedback (Joyce & Showers, 1981).

Teachers also stated that they wished the MPAC-T measure itself—not just the rater training program—included more detail about items. In particular they requested more specific behavioral indicators for items. This sentiment is supported in the empirical literature; teachers’ ratings on a well-operationalized rating scale are less likely to exhibit a negative halo effect than their ratings on a more abstract scale (Stevens, Quittner, & Abikoff, 1998). Where the MPAC-T measure can be revised to provide concrete behavioral indicators, such as specific definitions of turn-taking (i.e., each turn has a clear start and end) and sharing (i.e., giving some or all of a limited-quantity item), the accuracy of teachers’ ratings may improve.

2. **Provide More Focused Guidance on Items that Were Especially Difficult for Teachers to Rate Accurately.**

Quantitative analyses (RQ5) revealed six items on which teachers’ ratings were least accurate. It is not surprising that teachers’ agreement with experts varied by item; teacher-observer agreement on the Social Skills Rating Scales, for example, varies in significance depending on the factor (Elliot, Gresham, Freeman, & McCloskey, 1988). Teachers’ focus group responses (RQ7) highlight specific ways in which the MPAC-T rater training program might be revised to address misunderstandings about these six low-accuracy items:

- **Positive affect** (Items 1 and 1a): Teachers described inferring positive affect when children’s expression was neutral, for example assuming that a child felt happy but did not express it for cultural reasons. Thus the rater training program should encourage teachers to “check” only visibly expressed positive emotion, rather than make inferences.
• **Sharing and turn-taking** (Items 13 and 15, respectively): Teachers explained that they use the terms “sharing” and “turn-taking” interchangeably to refer to mutual use of an item. The MPAC-T rater training program should provide specific behavioral indicators to distinguish each item.

• **Fidgety and uninterested** (Item 6): Teachers described confusion when this item applied only partially. For example, a child might fidget but remain interested, or be disinterested in his own work but quite interested in someone else’s. The MPAC-T rater training program should specify the appropriate object of the child’s interest and the degree to which both fidgety behavior and disinterest are necessary to merit a “check.”

• **Involved in a self-organized activity** (Item 4): Teachers did not mention this item as particularly difficult to rate, yet they under-rated it as compared with experts (RQ5). This could suggest that teachers were unaware of Item 4; they did not think to rate it or to mention it during the focus group. So that this item is not overlooked in the future, the rater training program should more clearly emphasize Item 4’s unique emphasis on self-organized activity. The item wording, too, could be revised to highlight this: for example, “Helped to organize an activity and became actively involved in it.”

Teachers tended to over-rate these six low-accuracy items; they “checked” these items when the experts did not (see Figure 2). Similarly, a prior study found that novice coders tend to overestimate both positive and negative behaviors in children (Riley-Tillman et al., 2010).

Defining MPAC-T items more specifically, both in the rater training program and on the measure itself, may narrow the scope of behaviors that teachers feel it is appropriate to “check” for each item. Future iterations of the MPAC-T rater training program could also emphasize that
some observations will include fewer MPAC-T behaviors than others; there is no need to correct for this by “checking” additional behaviors.

3. **Shorten the Reliability Testing Phase**

   The idea that raters’ ratings may change over time is well established (Wolfe, Moulder, & Myford, 2001). For example, a prior study found that grading order significantly predicted the grades that teachers assign (Klein & Pat El, 2003). In the present study raters’ ratings became less accurate over the final three videos. This pattern may be attributable to fatigue, which decreases attentional resources over the course of the rating session (Wolfe et al., 2001).

   Teachers’ focus group responses (RQ7) support this finding, with approximately one quarter of participants mentioning that the post-test phase was too long, and no participants describing its length favorably. Future studies should include fewer or shorter post-test videos. When used in schools the MPAC-T should include some post-training videos to establish accuracy; however it would be wise to use fewer or shorter videos than in the present study.

4. **Correct Technical Glitches in the MPAC-T Rater Training Program**

   Approximately one-third of teachers reported technical glitches in the MPAC-T rater training program (RQ8)—most commonly freezing and restarting of the training videos (i.e., explanations with examples). Technical glitches are known drawbacks of online rater training programs (Australian National Testing Authority, 1998; Hamilton et al., 2001). Future iterations of the MPAC-T rater training program should use a more reliable video hosting service to reduce playback issues.
C. Factors that Could Affect the Accuracy of Teachers’ MPAC-T Ratings

1. Assistant versus Lead Teachers.

Several findings suggest that it may be wise to train assistant teachers, rather than lead teachers, to administer the MPAC-T. Teachers who were more educated and those who used assessment tools more frequently tended to rate children less accurately on the MPAC-T (RQ9). These teachers may have developed a fixed mindset about child behavior through their prior education and assessment use. This mindset may have made it difficult for them to internalize new MPAC-T conventions and definitions. Because assistant teachers may be less educated and have less experience using assessment tools, they may be more open to internalizing MPAC-T standards. Further, teachers indicated that they would find it difficult to administer the MPAC-T due to competing classroom demands (RQ10). Assistant teachers may have fewer responsibilities for planning and leading lessons; thus they may have more flexibility to administer the MPAC-T.

2. Knowledge of the Child

Teachers also described how their prior knowledge of the focal child would affect their MPAC-T ratings (RQ10). They believed that prior knowledge of the child would improve the accuracy of their ratings, providing more information to aid in their judgment. However, the converse could also be true; prior knowledge could lead teachers to make biased inferences that are not supported by observable evidence. This seems likely, given prior findings that teachers’ ratings can be biased by a child’s background. For example, a recent review found strong evidence of bias in teacher ratings when students violated teachers’ positive cultural stereotypes (Mason, Gunersel, & Ney, 2014). Preconceived notions about a child’s abilities or tendencies may lead to biased inferences on the MPAC-T.
D. Future Directions

Although these findings are promising, there are several remaining questions to be addressed before the MPAC-T can be successfully implemented in classrooms. One important future study concerns classroom applications of the MPAC-T. That is, how can MPAC-T ratings be used effectively for instructional purposes? Teacher focus groups should be convened to discuss potential uses for MPAC-T scores and ways that MPAC-T results can be presented to facilitate these uses. For example, teachers may wish to use MPAC-T scores to document individual students’ progress toward state learning standards. For this purpose, scores on individual MPAC-T items might be aggregated according to the learning standards they address. Alternately, teachers might be interested in pinpointing specific skills to reinforce with particular children; in this case reporting individual item scores for each child would provide the most precise information. Teachers might also use MPAC-T scores to gauge the social-emotional strengths and weaknesses of their classroom as a whole; for this purpose scores would be reported at the classroom level, rather than the individual level. Regardless of how MPAC-T scores are used, it will be important to specify that the MPAC-T is intended to guide instructional decision-making (e.g., which SECs to emphasize) rather than high-stakes decision-making (e.g., a child’s kindergarten readiness or a teacher’s effectiveness). Future teacher focus groups should also solicit teachers’ feedback on logistical aspects of MPAC-T administration, including decisions about who and when to assess. It may be most practical for teachers to focus on assessing only a select group of children, but whether those children are identified by the teacher or randomly selected by the program is yet to be determined.

Another important question is whether teachers can provide reliable, accurate MPAC-T ratings when live-rating children in the classrooms where they teach. Teachers’ responses
(RQ10) suggested that live-rating might differ from in-lab rating due to competing attentional demands and prior knowledge of the focal child. To investigate these and other differences, a trained teacher and an external observer could dual-rate children live in the teacher’s classroom. Ideally this dual-rating process would be repeated weekly on the same children over four weeks, as the MPAC-T is designed to be used. This would allow for an analysis of training fadeout effects; are teachers’ ratings less likely to agree with observers’ ratings at the fourth observation than at the first? Such a study could also explore differences in teacher accuracy by focal child culture, as teachers mentioned in focus groups that they considered culture when making ratings.

Another study could explore whether teachers alter their instructional practice as a result of rating their students on the MPAC-T, without any further intervention. For example, a controlled study found that teachers are more likely to provide behavioral prompts to students who they have observed systematically, and observed students show greater behavioral changes (Hay et al., 1977). Simply rating students using the MPAC-T may orient teachers to students’ strengths and weaknesses, altering their social-emotional teaching practices.

E. Limitations

Teachers completed the MPAC-T rater training program and post-test in the laboratory, which limits the external generalizability of my findings. Ideally participants would complete the rater training program at home and the post-test in their classroom, as these elements will ultimately be administered. However raters must be motivated to complete training and make ratings independently. In most studies participants are motivated to complete at-home rater training because it is required for their occupation (e.g., Knoch et al., 2007; Lundh et al., 2012). However the MPAC-T and rater training program are not yet developed enough for use in classrooms. Thus participants would have little internal motivation to complete training and
ratings on their own time, as evidenced by low video viewing duration during the at-home follow-up phase of this study. Thus, to demonstrate the efficacy of the MPAC-T rater training program under ideal circumstances, I chose for raters to complete training and ratings in the laboratory, much as other rater training studies have done (Chafouleas et al., 2012; Eschen-Leguede, & Peccena, 2006; Kobak et al., 2006; Rosen et al., 2008). Future studies should investigate the effectiveness of rater training conducted at home and MPAC-T ratings assigned in classrooms.

Without a control group it is difficult to attribute the pre-post improvement in accuracy solely to the rater training program. Practice effects or having a pre-test and post-test that were not equivalent could also explain this change. Future studies could employ counterbalancing of the pre-test and post-test videos or repeat one or more videos at pre- and post-test.

At follow-up teachers’ accuracy levels returned to baseline, indicating that training effects did not last. This finding may not be wholly attributable to fadeout. Teachers completed follow-up ratings at home, rather than in the lab; many of their viewing times, while not low enough to warrant exclusion, indicated a rushed approach to follow-up ratings. Thus the transition from laboratory to home and attendant decrease in attention may partially explain the decline in accuracy at follow-up. To reduce fadeout in the future, the MPAC-T measure itself could be revised to include more detail about each item. Teachers indicated a need for this in focus groups, stating that that the MPAC-T measure was less detailed than the training program. It is unlikely that accuracy would increase from post-test to follow-up, as teachers did not receive feedback on the accuracy of their post-test ratings.

The generalizability of my findings is limited in some ways by my sample. Teachers were excluded from this study if they taught exclusively in self-contained special education
classrooms; future studies could explore whether the MPAC-T and rater training program are appropriate for use in these contexts. Further, teacher subgroups by characteristics (i.e., assessment frequency, education, and teaching experience) were small, ranging in size from 2 to 12; it is important to replicate these findings in a larger sample before generalizing them to all early-childhood educators. Finally, my sample was both more educated and less experienced than the national average for center-based early-childhood educators (National Survey of Early Care and Education Project Team, 2013). Because teachers’ backgrounds may affect their ratings, a future study should investigate whether these findings generalize to teachers with different levels of experience and education, as well as those teaching in non-urban settings.

F. Conclusion

This study provides necessary formative feedback for the improvement of the MPAC-T rater training program. In this way it addresses a call for greater attention to rater training for teachers (Hamilton et al., 2001; Ledoux, Yoder, & Hanes, 2010). Further, it sheds needed light on teachers’ responses to rater training (Elder et al., 2007), indicating, for example, that their definitions of social-emotional terms may conflict with researchers’ definitions, as in the case of turn-taking and sharing. In testing a brief rater training program, the present study contributes to an important, emerging literature on the use of brief rater training to promote accurate, reliable observations of social behavior (Angkaw et al., 2006; Schlientz et al., 2009). Findings suggest that with 45 minutes of training, teachers can rate preschoolers’ social-emotional behaviors accurately and consistently. The MPAC-T empowers teachers to collect and use observations of students’ social-emotional behaviors; doing so may improve instructional practice and ultimately, students’ social-emotional competencies.
CITED LITERATURE


Minnesota Preschool Affect Checklist –Teacher (MPAC-T)

*Indicate if the target child did any of the following things at least once during the 5-minute observation. If a behavior has been checked off once, it does not need to be checked off again. To fully describe a single behavior, you may need to check off multiple items.*

**WHAT KINDS OF EMOTIONS DID THE CHILD EXPRESS?**
- □ 1. Showed **positive** emotions in any manner (i.e., facial, vocal, or bodily) that made sense given the situation.
  - □ 1a. The positive emotion was directed a particular person.
  - □ 1b. The positive emotion was **not** directed at a particular person.
- □ 2. Showed **negative** emotions in any manner (i.e., facial, vocal, or bodily affect) that made sense given the situation.
  - □ 2a. The negative emotion was directed at a specific person.

**HOW INVOLVED WAS THE CHILD IN CLASS ACTIVITIES?**
- □ 3. Was engaged and engrossed in an activity that has a positive emotional function.
- □ 4. Was involved in an activity that he/she helped to organize.
- □ 5. Looked “emotionally absent” and showed no involvement in an activity
- □ 6. Looked fidgety and uninterested or uninvested in the activity but still “tuned in” and aware of his/her surroundings.

**WHEN FRUSTRATED, HOW DID THE CHILD REACT?**
- □ 7. Reacted to a frustrating situation by aggressively attacking another person either verbally or physically.
- □ 8. Took frustration out on an object.
- □ 9. Reacted to frustration by talking about the problem and **did not stay upset**.
  - □ 9a. In his/her reaction to frustration, the child was neutral or positive.

**HOW DID THE CHILD INTERACT WITH OTHERS?**
- □ 10. Physically attacked another child for no apparent reason.
- □ 12. The child displayed leadership skills by organizing an activity in which another child or children “followed the lead” and participated.
- □ 14. Cooperated with peers to achieve a common goal.
- □ 15. Shared toys or other materials.
- □ 16. Engaged in no social interactions for 3 minutes or more.
Appendix B

Computerized Rater Training Outline
Minnesota Preschool Affect Checklist-Teacher (MPAC-T)

1. Overview of the MPAC-T (voiceover with outline text)
   a. Purpose
      i. Definition: Social-emotional competencies
      ii. Importance of SECs and MPAC-T ratings
      iii. Classroom uses of MPAC-T
   b. Conventions
      i. 5-minute free-play observation on 4 occasions
      ii. Checklist format: yes/no
   c. Sections
      i. What kinds of emotions did you see the child express?
      ii. How involved was the child in class activities?
      iii. When frustrated, how did the child react?
      iv. How did the child interact with others?

2. For each section (listed under 1c)
   a. Item-by-item explanation, presented via voiceover while the individual item is displayed in verbatim text. For each item:
      i. Voiceover reads each item aloud.
      ii. Voiceover describes behaviors that merit a “check.”
      iii. Brief video example of behavior that merits a “check.”
   b. Practice ratings, presented with video on one half of screen and rating items on the other half.
      i. Display 3-minute video clip containing some behaviors from this section.
      ii. While clip plays, users rate it using only the specified section of the MPAC-T.
   c. Feedback
      i. Visually indicate on which items user was correct or incorrect.
      ii. For each item the participant rated incorrectly:
         1. Display a written rationale for the correct rating (no voiceover).
         2. Provide an optional link to re-watch the practice video.

3. Whole-scale use (after the participant has completed training for all four sections)
   a. Practice ratings, presented with video on one half of screen and the MPAC-T on the other half.
      i. Play a 5-minute video that exemplifies some MPAC-T behaviors
      ii. While clip plays, users rate the clip using the entire MPAC-T (all sections).
   b. Feedback
      i. Visually indicate on which items user was correct or incorrect
      ii. For each item the participant rated incorrectly:
         1. Display a written rationale for the correct rating.
         2. Provide an optional link to re-watch the practice video.
Appendix C

Screenshots from the MPAC-T Pre-Test, Rater Training Program, and Post-Test

Image 1: Pretest Interface

Video 1 of 3: Navy Hat

Indicate if the target child did any of the following at least once during the 5-minute observation. If a behavior has been checked off once, it does not need to be checked off again. To fully describe a single behavior, you may need to check off multiple items.

Enter your name and click submit. Then play the video and code it. If the video does not load, try this website in Internet Explorer. If you are logged into the Box.com website, log out.

What kinds of emotions did you see the child express?

- 1. Shown positive emotions in any manner (e.g., facial, vocal, or bodily affect) that make sense given the situation.
  - 1a. The positive emotion was directed at a particular person.
  - 1b. The positive emotion was not directed at a particular person.
- 2. Shown negative emotions in any manner (e.g., facial, vocal, or bodily affect) that make sense given the situation.
  - 2a. The negative emotion is directed at a specific person.

How involved was the child in class activities?

- 3. Engaged and engrossed in an activity that has a positive emotional function.
- 4. Was involved in an activity that he/she helped to organize.
- 5. Looked ‘emotionally absent’ and showed no involvement in an activity.
- 6. Looked fitgally and uninterested or uninvolved in the activity but still ‘tuned in’ and aware of his/her surroundings.

The MPAC-T Pretest Interface is available at https://sites.google.com/a/psch.uic.edu/mpact/
The MPAC-T rater training program is available at
https://uic.qualtrics.com/SE/?SID=SV_0PnnCDgjPdFLTMe
Image 3: Practice Page from the MPAC-T Rater Training Program

Focal child: Navy hat

**What kinds of emotions did you see the child express?**

- □ 1. Showed **positive** emotions in any manner (i.e., facial, vocal, or bodily) that made sense given the situation.
  - 1a. The positive emotion was directed at a particular person.
  - 1b. The positive emotion was not directed at a particular person.
- □ 2. Showed **negative** emotions in any manner (i.e., facial, vocal, or bodily) that make sense given the situation.
  - 2a. The negative emotion is directed at a specific person.
Image 4: Feedback Page from the MPAC-T Rater Training Program

<table>
<thead>
<tr>
<th>Here are the correct responses for this video:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Showed positive emotions in any manner (i.e., facial, vocal, or bodily) that made sense given the situation.</td>
</tr>
<tr>
<td>1b. The positive emotion was not directed at a particular person.</td>
</tr>
</tbody>
</table>

One or more of your responses was incorrect.

Below you can read feedback on your work and re-watch the example video.

You did not select "1. Showed positive emotions in any manner (i.e., facial, vocal, or bodily) that make sense given the situation."

This item does apply to the video. The child begins smiling around 0:12.

You did not select, "1b. The positive emotion was not directed at a particular person."

This item does apply to the video. When the child smiles (around 0:12) he is not looking at any particular person.

You selected "2a. The negative emotion was directed at a specific person."

This item does not apply to the video. Although the child speaks authoritatively at the beginning of this clip, it's not clear whether his emotions are negative.

Re-Watch the Video

Affect (Navy Hat).mp4
Image 5: Follow-Up Interface

The MPAC-T follow-up test website is available at https://sites.google.com/a/psch.uic.edu/mpac-t-follow-up/
Appendix D
Focus Group Guide
Minnesota Preschool Affect Checklist-Teacher
Rater Training Development

Focus Group Procedures
Before the focus group begins, the assistant moderator will set up the camera and refreshments. The lead moderator will welcome participants, offer them refreshments, and give them identification number signs (paper tents). Once participants are seated, the assistant moderator will draw a seating chart with identification numbers and check the camera angle. The lead moderator will use the following script to start the group:

Hello everyone. Thank you for agreeing to participate in today’s discussion. My name is Claire Christensen, and I am a graduate student at the University of Illinois at Chicago. Assisting me is __________, also from UIC. I’m going to briefly introduce what we will be doing, then we’ll introduce each other and finally I’ll start the video camera to record our discussion.

We asked you here because you are all early childhood educators and we are hoping that you can help us understand your reactions to the observational rating tool you just used, the Minnesota Preschool Affect Checklist –Teacher, or MPAC-T, and the training program you completed at the start of this session. I’m going to pose a few questions to you all over the next hour and I hope you will share as much as you are comfortable.

Before we start recording I want to remind you of a few points. First, neither my assistant nor I knows anything about how you did in the training program or how you rated the videos. Second, when we transcribe this focus group, any names will not be transcribed. If you use a child’s name or colleague’s name, don’t worry, we won’t be transcribing that. Third, to make it easier for us to transcribe this discussion, please try to avoid talking over each other. We’ll have time for everyone to contribute so let’s try to go one at a time. Finally, although we are recording video, we will analyze only the audio portion of this conversation. Don’t worry about how you look on camera. We are only using a video camera so that we can associate your anonymous ID number with the things you say.

Do I have everyone’s consent to begin video recording? If not, please raise your hand.

At this time the assistant moderator will turn on the video camera. From this point on the assistant moderator will check the camera periodically to ensure that it is still recording. The lead moderator will say:

Before I pose the first question, I’m going to ask each of you to read the number on the paper in front of you so we can match your voice with your identification number on the tape. We’ll start here (point to teacher on the left).

Immediately after the focus group, the lead moderator will thank participants for participating in the discussion and remind them about the follow-up component. The lead moderator will have
participants sign for cash compensation and double check their identification numbers against the seating chart.

The assistant moderator will break down equipment and refreshments. Back at the lab, the assistant moderator will upload the video and save it to the lab hard drive. She will also type up the focus group seating chart. The lead moderator will strip the video files of any identifying information prior to transcription.

**Semi-Structured Focus Group Script**

I’d like to talk about three aspects of your experience here today. First we’ll talk about the training program you used at the beginning of this session, which included the voiceover explanations, the short practice videos that you rated, and the feedback you received. Second we’ll talk about the twenty-item rating instrument that you used after you finished training. Finally we’ll talk about what it was like to rate those ten videos using the rating instrument.

1. First let’s talk just about the training program you used at the beginning of this session. (Assistant moderator hands out screenshots from the training program.) How would you describe your experience using the training program?
   a. What did you like about the training program? (Prompt: What about the training program was helpful or enjoyable?)
   b. How would you change the training program to improve your experience?
      i. What about the training program was difficult to use or understand?
         (Prompts: technical issues, navigational issues, personal preferences.
   c. Imagine you needed to complete the training program on your own, at home or at school. How would your training experience be different? (Prompts: Would you complete it in one sitting? Would you spend more or less time on it than you did today?)
      i. How could the training program be changed to help teachers complete it on their own?

2. After you completed the training program, you used the whole rating instrument, with all twenty items. (Assistant moderator hands out paper copies of the MAPC-T.) Let’s talk just about that rating instrument. What are your impressions of the rating instrument?
   a. What did you like about the rating instrument? (Was there anything about the instrument that made it easier to use?)
   b. What changes would you like to see in the rating instrument? (Prompts: What was difficult to understand? Which items were you unsure about?)

3. After the training program, you rated ten videos. What was it like to rate those videos? (Prompts: How difficult was it? What were you feeling? What were you thinking?)
   a. Which behaviors were easy to rate? (Prompt: What made it easy?)
   b. What about the rating process was difficult?
   c. What questions came up for you?
   d. Before you came here today, you rated three videos at home. Did the training change the way that you rated videos or how you felt about the task? How so?
      i. How could the training program have better prepared you to rate videos?
e. Imagine you were going to use the rating instrument to observe one student in your class for five minutes during free play. How would that experience be different from what you did today?
   i. How could the rating instrument be changed to help teachers use the rating instrument in their classrooms?
   ii. How could the training program better prepare teachers to rate their students?

4. Let me tell you a little more about why we’re so interested in hearing from you. We are developing this rating instrument and training program to help teachers measure their students’ social-emotional strengths and weaknesses so that they can decide how best to promote social-emotional skills in their classrooms. We want to understand how to improve the rating instrument and the training program to help teachers use them in their classrooms. With that in mind:
   a. What do you think is the most important thing for us to know about making a rating instrument and training program for teachers to use?
   b. Is there anything else we need to consider? Any last comments?
Appendix E

Multi-Facet Rasch Measurement Models

Intra-Rater Reliability Analysis

\[
\log \left[ \frac{P_{nijk}}{P_{nij(k-1)}} \right] = B_n - D_i - C_j - F_{ik} \tag{1}
\]

where

- \(P_{nijk}\) is the probability that video \(n\) will receive a rating of 1 from teacher \(j\) on item \(i\),
- \(P_{nij(k-1)}\) is the probability that video \(n\) will receive a rating of 0 from teacher \(j\) on item \(i\),
- \(B_n\) is the degree of social-emotional competency depicted in video \(n\),
- \(D_i\) is the difficulty of item \(i\),
- \(C_j\) is the severity of teacher \(j\), and
- \(F_{ik}\) is the difficulty of scale category \(k\), relative to scale category \(k-1\) for item \(i\).

Accuracy Analysis

\[
\log \left[ \frac{P_{nidjqk}}{P_{nidjq(k-1)}} \right] = B_n - D_i - C_j - K_d - L_q - F_{ik} \tag{2}
\]

where

- \(P_{nidjqk}\) is the probability that a teacher with teaching experience \(j\), education \(d\), and assessment use \(q\) will assign an accurate rating \((x = 1)\) on item \(i\) for behavior observed in video \(n\),
- \(P_{nidjq(k-1)}\) is the probability that a teacher with teaching experience \(j\), education \(d\), and assessment use \(q\) will assign an inaccurate rating \((x = 0)\) on item \(i\) for behavior observed in video \(n\),
- \(B_n\) is the ease of assigning accurate ratings for the behaviors observed in video \(n\),
- \(D_i\) is the ease of assigning accurate ratings on item \(i\),
- \(C_j\) is the accuracy of a teacher with teaching experience \(j\),
- \(K_d\) is the accuracy of a teacher with education \(d\),
L_q = the accuracy of a teacher with assessment use q (i.e., how frequently the teacher uses classroom assessments, classified as rarely, sometimes, all the time, or often),

F_{ik} = the difficulty of assigning an accurate rating in scale category k, relative to assigning an accurate rating in scale category k - 1 for item i.
SUMMARY

I have 6 years of experience in social science research, including quantitative and qualitative studies. My strengths include:

- Designing and leading site-based research and evaluation studies.
- Managing and analyzing large, complex datasets.
- Communicating research-based information through reports, presentations, and peer-reviewed publications.

EDUCATION

University of Illinois at Chicago
Ph.D., Psychology: Community and Prevention Research 2015
Minor: Educational Program Evaluation

University of Illinois at Chicago
M.A., Psychology 2011
Thesis: Measuring Social and Emotional Content in Educational Television for Children

Illinois State University
B.S., Mass Communication 2008
Minor: Psychology

RESEARCH & EVALUATION EXPERIENCE

Social Emotional Teaching & Learning Lab, University of Illinois at Chicago 2013 – Present
Laboratory Manager
- Co-authored a research study on the effectiveness of Head Start policies for reducing teacher stress.
- Managed a team of six researchers in a site-based evaluation including over 200 observations and assessments.
- Designed and led a year-long mixed-methods program evaluation, including focus groups with 22 educators.
- Used SPSS and NVivo to analyze focus groups, surveys, and structured observations.
- Collaborated on three grant proposals, two of which received funding.

Early Investments Initiative, Institute of Government and Public Affairs 2015 – Present
Research Assistant
- Planned and managed over 100 hours of data collection for an education policy evaluation.
- Co-authored a policy brief on state learning standards and quality rating and improvement systems.

Michael Cohen Group 2014 – 2015
Research Assistant
- Collaborated on two federally funded summative evaluations of PBS Kids educational programs.
- Administered over 100 structured interviews and STEM assessments for experimental program evaluations.

Museum of Science and Industry 2013 – 2014
Research Intern
- Collaborated on four program evaluations from planning through data collection and analysis.
- Used ATLAS.ti to analyze 41 qualitative interviews; produced reports about program effectiveness.
Center on Media and Child Health, Boston Children’s Hospital 2011 – 2014
Research Intern
• Analyzed a large health behaviors dataset; disseminated findings in a first-author peer-reviewed publication.

Department of Psychology, University of Illinois at Chicago 2010 – 2014
Teaching Assistant
• Planned and taught weekly classes in research methods, statistics, and developmental psychology.
• Advised 150 students on research design and statistical analyses for original research studies.

Social Emotional Learning Research Group, University of Illinois at Chicago 2009 – 2010
Research Assistant
• Analyzed program evaluation studies to determine methodological rigor and program effectiveness.
• Communicated program evaluation results in a written guide for educators.

Pearson Education 2007 – 2009
Research and Editorial Assistant
• Served as project manager for the development of educational games and videos.
• Wrote nationally published study guides and teaching supplements for collegiate courses.

GRANTS AND HONORS

• Provost’s Award ($1,400). University of Illinois at Chicago. 2014
• Chancellor’s Graduate Research Fellowship ($8,000). University of Illinois at Chicago. 2011—2013
• Suma Cum Laude. Illinois State University. 2008

PEER-REVIEWED PUBLICATIONS


MANUSCRIPTS IN PREPARATION


OTHER PUBLICATIONS


PRESENTATIONS OF RESEARCH


CONFERENCE POSTERS


**INVITED TALKS**


**PROFESSIONAL MEMBERSHIPS**

- American Educational Research Association
- Society for Research in Child Development
- American Evaluation Association
- American Psychological Association, Division 15: Educational Psychology
- Early Investments Consortium, Institute of Government and Public Affairs

**EDITORIAL EXPERIENCE**

- Child Development 2008-present
- Midwest ECO Conference 2013
- Early Education and Development 2013

**SKILLS**

- SPSS
- NVivo
- ATLAS.ti
- Qualtrics
- MS Office

**ASSESSMENT CERTIFICATIONS**

- Bracken School Readiness Assessment
- Affect Knowledge Test
- Challenging Situations Task
- Minnesota Preschool Affect Checklist
- Preschool Classroom Observation Record