Observing Classroom Practices with the Classroom Assessment Scoring System – Secondary (CLASS-S): A Scoping Review, and Preliminary Reliability and Validity Study in Alberta, Canada

by

Stephanie Anne Hayes

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

SCHOOL AND CLINICAL CHILD PSYCHOLOGY

Department of Educational Psychology
University of Alberta

Abstract

Delivering quality instruction that fosters student learning is a complex and challenging task for teachers, particularly given the evolving needs of 21st century education (Downer, Jamil, Maier, & Pianta, 2012; McCombs, 2013). Despite a long tradition of classroom observations in educational research, researchers have only recently given more attention to standardized observational systems of classroom quality. In particular, researchers have recognized standardized observations might acknowledge the complexity of classroom practices, allow for contrasts across diverse educational settings, and support teacher development and accountability initiatives (Cantrell & Kane, 2013). One such observational protocol, the Classroom Assessment Scoring System – Secondary (CLASS-S), has been used as an observational tool in a variety of research examining quality classroom practices in middle and high schools in the United States (Pianta, Hamre, & Mintz, 2012). Emerging support for its use outside of the United States has also been published in the United Kingdom (Malmberg, Hagger, Burn, Mutton, & Colls, 2010), Finland (Virtanen, Pakarinen, Lerkkanen, Poikkeus, Siekkinen, & Nurmi, 2018), and Norway (Westergård, Ertesvåg, & Rafaelsen, 2018). Researchers have found evidence that higher quality classroom practices, as measured by the CLASS-S, are associated with positive student outcomes (e.g., Allen, Pianta, Gregory, Mikami, & Lun, 2011; Culp, Martin, Clements, & Presser, 2015; Kane & Staiger, 2012; Virtanen, Lerkkanen, Poikkeus, & Kuorelahti, 2015). The present research

had two aims: 1) to provide a systematic scoping review of the use of the CLASS-S in internationally published research in middle and high school classrooms and; 2) to explore the reliability and validity of the CLASS-S, drawing from a sample of 39 middle school teachers who participated in a larger research project across five different educational jurisdictions in Alberta, Canada. In brief, the scoping review examined the use of the CLASS-S in published literature, mapped key themes addressed by authors, presented findings about the relationship between the CLASS-S and student outcomes, and identified gaps and summarized recommendations for the future use of the CLASS-S. Despite identifying more research is needed to support the use of CLASS-S; this tool has the potential to enrich our knowledge of educational practices, support teachers' classroom practices, and enhance teacher-student interactions in middle and high school settings. Furthermore, findings from the preliminary reliability and validity study of the CLASS-S identified emerging support for its use in Alberta, Canada. Finally, recommendations for further research to deepen the emerging evidence-base and contribute to our understanding of the complexities of classroom practices as captured by the CLASS-S are summarized.

Preface

Some of the research conducted for this dissertation was part of a larger research project, entitled "Research project to gain a deep understanding of promising practices, implementation strategies and educational benefits/results of technology use within inclusive Junior High/Middle School settings", led by Dr. Veronica Smith at the University of Alberta. However, the papers presented here are original work by Stephanie A. Hayes as lead author, and received ethics approval as a secondary analysis of the data from the larger project from the University of Alberta Research Ethics Board, Project Name "Observing Teacher-Student Interactions Using the Classroom Assessment Scoring System – Secondary (CLASS-S) in Alberta", Study ID: Pro00066357, July 26, 2016 (see Appendix A). In addition, a graduate research assistant, Krystle-Lee Turgeon, contributed to the scoping review presented in Chapter Two as the second coder. Dr. Veronica Smith provided ongoing guidance and feedback in the preparation of this manuscript.

Dedication

In loving memory of my mother-in-law, Armi Nisbett. Your strength and love supported our family, and provided me with opportunities for growth. Your time with Rylann not only allowed me to realize this dissertation, but also gave our daughter a beautiful foundation of love and friendship. Always in our hearts – forever in the stars.

Acknowledgement

I wish to acknowledge with deep gratitude those who have contributed, shaped, and supported my growth. First and foremost, thank you to my mentor, Dr. Veronica Smith, for providing me with opportunities, guidance, encouragement, and friendship. Dr. Smith's confidence in my abilities as a researcher, teacher, and clinician has helped me cultivate a professional sense of self. Thank you to my committee members, Dr. Jacqueline Pei and Dr. George Buck for guiding me through my graduate training and the dissertation process. I am also appreciative to Dr. Pei as a mentor, for the experience gained as a research assistant on the Professionals without Parachutes project. I would also like to thank Drs. Cutumisu and Bulut for their feedback during the final stages of preparation of this manuscript.

I would like to thank my partner Scott for his support, patience, and willingness to begin new adventures like moving across the country, traveling the world, and starting a family. Thank you to my daughter, Rylann, for showing me the importance of slowing down and being in the moment; and my new baby boy, Dax, who arrived just as I was completing this manuscript, for completing our family. Thank you to my mother Vicki and sister Jillian, for their enduring encouragement, for listening when I was frustrated, and for celebrating my accomplishments.

Finally, I am grateful to the many people I have crossed paths with on this journey who have contributed to my learning and growth, including friends, colleagues, professors, clinical

supervisors, families, teachers, administrators, and students. In particular, Krystle-Lee Turgeon, who was always there to provide support as a dear friend, and came through as a hard-working graduate research assistant on the Flexible Pathways project and secondary coder for part of this dissertation. I would like to give a special thanks to my colleagues at Algoma Family Services, especially Dr. Carol Fick, Roy Thompson, Gina Creighton, and Sandie Leith, for providing me with a rich internship experience, and continued encouragement in the final phase of my graduate training.

I would also like to acknowledge the funding support I have received over the course of my doctoral training including; the Joseph-Armand Bombardier Canada Graduate Scholarship – Social Sciences and Humanities Research Council of Canada, the Queen Elizabeth II Graduate Scholarship, the University of Alberta President's Doctoral Prize of Distinction, the University of Alberta Recruitment Scholarship, the Canadian Institute of Health Research Strategic Training Initiative in Health Research Autism Research Training (ART) program, and the Northern Ontario Internship Grant.

Table of Contents

List of Tables

Chapter Two.

Chapter Three.

List of Figures

Chapter Two.

Chapter Three.

List of Appendices

Chapter One. Introduction

Engaging in the practice of classroom teaching to foster positive student outcomes is both complex and challenging (Downer, Jamil, Maier, & Pianta, 2012). Furthermore, it has been widely recognized that educational transformation must occur in order to meet the demands of the 21st century (e.g., Government of Alberta, 2010; McCombs, 2013). Whereas traditional education emphasized learning content, in a world where information is now easily accessed, 21st century skills emphasize a student's ability to evaluate, analyze, and apply knowledge to solve novel problems (O'Sullivan & Dallas, 2017). Therefore, reform efforts have focused on the need for improvements in instruction, professional development, and increased accountability (McCombs, 2013). Moreover, researchers have been advocating for research-validated frameworks drawing on what is known about theories of learning, motivation, and development, to inform educational transformation (McCombs, 2013). For example, in 2010 the Government of Alberta recognized that, despite having a world-class education system, classrooms in Alberta must transform to meet future needs (Government of Alberta, 2010). Therefore, classroom practices are increasingly under scrutiny, as policymakers press for change.

Importantly, traditional concepts of effective classroom practices and measures of student achievement have frequently been founded on standardized tests developed in the 20th century, based on 19th century principles of education (Berry, 2011). Therefore traditional concepts rely

on methods originally created for a different educational era, and draw on outdated tests, statistical procedures, and educational values and theories. Thus, new methodologies are needed to reflect our evolving conceptualization of classroom practices, and to deepen our understanding of how to support teachers and students to meet new expectations (McCombs, 2012). One important method commonly used to deepen our understanding of the complexities of teaching is classroom observations.

Classroom observations have been vital to understanding educational practices; however, traditionally, classroom observations have been unstandardized, informal, or based on locally developed or un-validated protocols (Stuhlman, Hamre, Downer, & Pianta, 2010). Despite the use of classroom observations in over thirty years of educational research, it is only recently that researchers have sought to develop valid and reliable standardized classroom observational protocols (Pianta & Hamre, 2009). Fuelled by innovations in technology that facilitate the use of video recording of classroom interactions (Klette & Blikstad-Balas, 2018), classroom observations have become increasingly more common, and a variety of measurement tools have been developed to target specific settings or subject areas (e.g., Harms & Clifford, 1998; National Center for Teacher Effectiveness, 2012; Stanford University, 2013), or to capture global practices in the classroom (e.g., Danielson, 2011; Pianta, La Paro, & Hamre, 2008). Video recordings have provided opportunities for researchers to thoroughly review and analyze

classroom practices, and systematically review the psychometric properties and utility of various classroom observational protocols (Klette & Blikstad-Balas, 2018).

Classroom Observational Protocols

In brief, classroom-based systematic observational protocols first emerged as primarily frequency or presence/absence checklist tools, where observers would code the occurrence of a given behaviour within a specified block of time (e.g., the Flanders Interaction Analysis Categories, the Stallings Observation System; Freiberg & Waxman, 1988). These early measures may have oversimplified classroom interactions by relying on the observation of specific components of teacher behaviour (e.g., reinforcement, reading training, cues, and feedback; Seidel & Shavelson, 2007) and this approach likely lacks the sensitivity to capture the larger scope and quality of classroom practices. In response to these shortcomings, the developers of modern observational protocols have sought to improve reliability and validity in their creation of more comprehensive measures (Pianta & Hamre, 2009). Furthermore, researchers have worked to make links between classroom practices and their associations with student outcomes (e.g., student achievement).

Of the myriad of observational measures that have emerged from the research, two general pedagogical instruments, the Framework for Teaching (FFT; Danielson, 2011) and the Classroom Assessment Scoring System (CLASS: e.g., Pianta, Belsky, Vandergrift, Houts, &

Morrison, 2008) have been widely adopted. Researchers compared the underlying constructs of the FFT and CLASS and identified they capture similar classroom practices (Gill, Shoji, Coen, & Place, 2016; Klette & Blikstad-Balas, 2018). Specifically, researchers' performed a content analysis on the FFT and CLASS and identified eight dimensions in common across both protocols. These dimensions were labeled as: supportive learning environment, student focus, classroom management, student intellectual engagement with content, lesson structure and facilitation, content understanding, language and discourse, and feedback and assessment (Gill et al., 2016). Furthermore, the researchers' identified two dimensions that differed between the FFT and the CLASS; the FFT does not include an indicator of students' active participation, and the CLASS does not include an indicator of teacher professionalism (Gill et al., 2016).

Drawing on knowledge from both practice and research, Danielson developed the FFT with the intent that its use would guide self-assessment and reflection, assist in pre-service teacher education, guide peer coaching, support supervision and evaluation, and enhance professional development (Danielson, 2011). The FFT includes four broad domains (planning and preparation, classroom environment, instruction, and professional responsibilities) and comprises 22 components (with 76 further smaller elements; Danielson, 2011). Classroom practices are rated on a four-point scale (i.e., unsatisfactory, basic, proficient, or distinguished; Danielson, 2011). Further, the use of the FFT does not require any specific training or

certification (Klette & Blikstad-Balas, 2018). Despite adoption across school jurisdictions as a tool for the evaluation of teachers, there appears to be limited published research exploring either the use of the FFT or its psychometric integrity (Pianta & Hamre, 2009).

In contrast, the CLASS has been used extensively in published research. The CLASS is based on the Teaching Through Interactions Framework (TTIF), a model supported by developmental theory and research evidence of effective instructional practices and teacher-student processes known to promote positive student outcomes (e.g., Battistich, Schaps, & Wilson, 2004; Emmer & Stough, 2001; Vermette, Foote, Bird, Mesibov, Harris-Ewing, & Battaglia, 2001; Zohar & Dori, 2003). The CLASS includes three broad domains of quality teaching (emotional support, classroom organization, and instructional support; Pianta & Hamre, 2009). Each domain includes three to four dimensions and trained, certified observers rate various behavioural indicators of teacher-student interactions for each dimension using a seven-point scale (Pianta & Hamre, 2009). Scores in the 1-2 point range are classified as low, in the 3-5 point range as mid, and in the 6-7 point range as high (Pianta, Hamre, & Mintz, 2012). The CLASS was designed to capture observations related to effective classroom practices while transcending boundaries related to content area or grades, as such, versions are available for use in all classrooms from pre-school through high school (e.g., Pianta & Hamre, 2009, Pianta, Hamre, & Mintz, 2012). Importantly, although the three broad domains remain consistent across

grade levels, the dimensions and behavioural indicators vary somewhat based on the developmental stage of the grade being observed (Pianta & Hamre, 2009).

What Can Observational Protocols Contribute to Our Understanding of Classrooms?

In developing standardized methods to capture quality instruction, researchers have sought to bridge the research-to-practice gap by linking observed practices to student outcomes (e.g., student achievement, improved behaviour). Beyond a student's own ability, Hattie (2003) proposed that teachers contribute 30% of the variance to academic achievement, and suggested quality teaching was the single most powerful influence on students. Therefore, observational protocols may provide valuable information linking quality classroom practices to student outcomes.

In a large-scale research project that included over 3000 classrooms, researchers explored the use of various observational measures (including the FFT and CLASS) and found that higher scores were associated with better student achievement, across all instruments (Kane & Staiger, 2012). Student achievement was measured in three ways; using state administered tests, and two supplemental assessments addressing math and reading (Kane & Staiger, 2012). The researchers concluded that combining observational measures with other more traditional measures of effective classroom practices enhances our ability to identify quality teachers beyond typical indicators such as teacher credentials and years of experience (Clotfelter, Ladd, & Vigdor, 2006;

Kane & Staiger, 2012; Wayne & Youngs, 2003). Furthermore, beyond student achievement, the students of quality teachers reported exerting more effort and experiencing more enjoyment in their classrooms (Kane & Staiger, 2012).

Researchers using the CLASS across a variety of settings have explored associations between the domains of the observational measures and various student outcomes. For example, researchers examining 224 prekindergarten classrooms across the U.S. found that CLASS domain scores related to the total score on the Early Childhood Environmental Rating Scale (ECERS), a standardized tool used to assess the quality of childcare settings associated with positive child development (La Paro, Pianta, & Stuhlman, 2004). A similar study used the CLASS and the ECERS to explore the quality and characteristics of prekindergarten classrooms in Spain (Sandstorm, 2012). In a study of student outcomes in 49 kindergarten classrooms in Finland, researchers found that higher scores on the CLASS domain of classroom organization and lower teacher stress predicted higher student learning motivation, which in turn predicted greater levels of student phonological awareness (Pakarinen, Kiuru, Lerkkanen, Poikkeus, Siekkinen, & Nurmi, 2010). Furthermore, in a large U.S. study comparing low-risk and at-risk first-grade students, researchers found that at-risk students placed in classrooms with higher scores in the emotional support and instructional support domains were similar to their low-risk peers with regards to student achievement (Hamre & Pianta, 2005). While in elementary school,

higher levels of emotional support contributed to a model of growth in math and reading achievement (Pianta, Belsky, et al., 2008). Finally, across 37 secondary school classrooms, researchers identified that, according to the CLASS, classrooms with higher scores on specific dimensions (i.e., positive climate, regard for adolescent perspectives, instructional learning formats, and analysis and inquiry) were predictive of greater student achievement, even when baseline achievement and other factors were controlled (Allen, Gregory, Mikami, Lun, Hamre, & Pianta, 2013). In sum, these examples suggest the CLASS is an observational measure associated with a variety of student outcomes related to learning and achievement.

Classroom Observations and Secondary School Classrooms

Since its development, the CLASS has been used across a breadth of published research reports, academic articles (i.e., peer-reviewed) and unpublished dissertations. However, much of the focus has been on its use in early years and elementary education. In contrast, the use of the Classroom Assessment Scoring System – Secondary (CLASS-S) version designed for middle and high school classrooms has received less attention. One of the underlying assumptions of the TTIF is that learning and development are fundamentally a function of ongoing classroom interactions between adolescents, adults, and peers (Hafen, Hamre, Allen, Bell, Gitomer, & Pianta, 2015). Furthermore, the CLASS-S takes into account other adolescent theories of development, such as the need for autonomy, competence, and relatedness in the classroom

(Ryan & Deci, 2000; Hafen et al., 2015). Therefore, the CLASS-S represents a shift in the conceptualization of quality middle and secondary classroom practices, by emphasizing theories of adolescent development over the more traditional emphasis on content delivery (Hafen et al., 2015). Further, much research on secondary classrooms have focused on student achievement (e.g., grade point average, results on standardized tests) as predictors of high school completion and educational attainment; however, researchers have been increasingly identifying that other, non-cognitive factors impact student performance (Farrington, Roderick, Allensworth, Nagaoka, Keyes, Johnson, & Beechum, 2012). Non-cognitive factors are those beyond a student's pre-existing cognitive ability. Researchers propose these factors are malleable and may lead to increased positive student outcomes, such as increased engagement in education (Farrington et al., 2012). Consequently, conducting classroom observations at the middle and secondary school level provides an opportunity to deepen our understanding of the non-cognitive processes and classroom practices that have potential to impact students beyond their pre-existing abilities. The CLASS-S developers have summarized links between theories of development and each dimension included in the observational protocol (e.g., attachment theory and Negative Climate, higher-order thinking and Analysis and Inquiry; Hafen et al., 2015).

Focus of the Present Research

As the body of research using the CLASS-S grows, now is the time to pause and

critically examine the evidence base of its use. The application of the CLASS-S as a standardized, reliable and valid observational protocol deserves attention in the current era where educational policy makers and administrators are seeking methods to be accountable and raise teaching standards in schools. Understanding how to define and identify quality classroom practices, and their relationship to student outcomes, provides policy makers and administrators with a benchmark for standards, and goals for improvement. How can we support teachers in their complex and challenging endeavours in the classroom, if we cannot reliably and accurately identify effective, quality classroom practices? According to Klette and Bilkstad-Balas (2018), the use of standardized observation protocols may facilitate an ongoing dialogue of what constitutes quality classroom practices by establishing a common framework and language, and by allowing for comparisons across settings. The authors noted that teaching has been typically described as a lonely profession, whereby knowledge about teaching was often based on individual "tinkering" and experience (Klette & Bilkstad-Balas, 2018). In contrast, the use of systematic observational protocols, such as the CLASS-S, may have implications for generating new knowledge about quality classroom practices and methods to support teachers in the classroom (Klette & Bilkstad-Balas, 2018).

Context and Overview

As previously noted, the Government of Alberta (2010) has called for educational

transformation to support 21st century skills in the classroom. To further support educational transformation, the province of Alberta developed The Learning and Technology Policy Framework (Alberta Education, 2013) and subsequently funded technology-based research projects aligned with the framework. The present dissertation is a secondary analysis of data gathered for one such provincially funded research project, "Flexible Pathways to Success: Technology to Design for Diversity" (Flexible Pathways Project; Smith, Hayes, Labonté, & Vargas, 2016). The Flexible Pathways Project was a two-year collaborative project that included Alberta Education, five unique school jurisdictions, and researchers from the University of Alberta. The goal of the project was to support the implementation of technology-based initiatives (as defined by each of the independent school jurisdictions) to support diverse learning needs in junior high or middle school classrooms. Drawing on a developmental evaluation approach (Patton, 2011), the researchers gathered mixed-methods summative and formative data (e.g., interviews, focus groups, questionnaires, video recorded classroom lessons) from various sources (e.g., administrators, teachers, students). The research framework was informed by several theories relevant to educational transformation and 21st century learning, including the Substitution, Modification, and Redefinition Model (e.g., Puentadura, 2010), Universal Design for Learning (e.g., Rose & Meyer, 2006), and the TTIF (e.g., Hafen et al., 2015). Findings from administrative, classroom, and individual student levels were shared with collaborators (i.e.,

Alberta Education, administrators, teachers) on an ongoing basis throughout the project.

The present dissertation focuses on the TTIF, and more specifically the CLASS-S (Pianta et al., 2012). What follows includes two articles (Chapter Two and Three) and an over-arching discussion and conclusion (Chapter Four). Chapter Two is a systematic scoping review of the use of the CLASS-S in the literature, while Chapter Three is a preliminary exploration of the reliability and validity of the CLASS-S in a sample obtained from schools across Alberta that participated in the Flexible Pathways Project. The aims of the scoping review (Chapter Two) included: (1) Provide an overview of the CLASS-S in published literature (e.g., who is using it, how are they using it?), (2) Map key themes addressed by authors using the CLASS-S, (3) Present findings describing relationships between classroom (i.e., CLASS-S) to student outcomes, and (4) Identify gaps in the literature and summarize recommendations for the use of the CLASS-S. The aims of the reliability and validity study (Chapter Three) include: (1) Report on the reliability (i.e., the item and scale reliabilities, and IRR of the CLASS-S), and (2) structural validity (i.e., the factor structure of the CLASS-S), of the observations conducted as part of the Flexible Pathways Project, across five unique jurisdictions in Alberta.

In order to deepen our understanding of classroom practices, and support educational transformation, it is first imperative we look at our commonly used methods with a critical lens. Classroom observations, and the CLASS-S in particular, have the potential to deepen our

understanding of classroom interactions. Of the protocols available, the CLASS-S has been used as a valid and reliable standardized protocol by researchers in the United States (e.g., Hafen et al., 2015; Kane & Staiger, 2012), Finland (Virtanen, Pakarinen, Lerkkanen, Poikkeus, Siekkinen, & Nurmi, 2018), and Norway (Westergård, Ertesvåg, & Rafaelsen, 2018). Importantly, the CLASS-S has not yet been used in published research in Canada. As such, pausing to review what is known about the CLASS-S internationally, and exploring its preliminary validity and reliability in an Albertan context, may establish the CLASS-S as a potential tool to support educational transformation in Canada.

Finally, a general discussion including a thematic summary of findings drawn from both studies presented in this dissertation will be provided (Chapter Four). Specifically, Chapter Four identifies themes related to the psychometric properties of the CLASS-S, associations between the CLASS-S and student outcomes, and the emerging use of the CLASS-S in international research, as particularly salient across the present research. What was learned and what remains to be learned will be reviewed.

References

Alberta Education. (2013). *Learning and technology policy framework.* Retrieved from https://education.alberta.ca/media/1046/learning-and-technology-policy-framework-web.pdf

Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B., & Pianta, R. (2013). Observations of Effective Teacher-Student Interactions in Secondary School Classrooms: Predicting Student Achievement With the Classroom Assessment Scoring System--Secondary. *School Psychology Review, 42*(1), 76-98.

Battistich, V., Schaps, E., & Wilson, N. (2004). Effects of an elementary school intervention on students' "connectedness" to school and social adjustment during middle school. *Journal of Primary Prevention, 24*, 243–262. doi:10.1023/B:JOPP.0000018048.38517.cd

Berry, B. (2011). Teacherpreneurs: A more powerful vision for the teaching profession. *Phi Delta Kappan, 92*(6), 28-33. Retrieved from https://wyoming2011uwyo.pbworks.com/f/Teacher%20Entrepreneurs%20Barnett.pdf

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources, 41*(4), 778-820. doi: 10.3368/jhr.XLI.4.778

Danielson, C. (2011). *Enhancing Professional Practice: A Framework for Teaching.* Alexandria,

VA: ASCD.

Downer, J. T., Jamil, F., Maier, M. F., & Pianta, R. C. (2012). Implications of information processing theory for professional development of early educators. In C. Howes, B.K. Hamre, & R. C. Pianta (Eds.) *Effective Early Childhood Professional Development: Improving Teacher Practice and Child Outcomes.* Baltimore, MD: Paul H. Brookes.

Emmer, E., & Stough, L. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist, 36*, 103–112. doi:10.1207/S15326985EP3602_5

Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., & Beechum, N. O. (2012). *Teaching adolescents to become learners. The role of noncognitive factors in shaping school performance: A critical literature review.* Chicago, IL: Consortium on Chicago School Research. Retrieved from https://files.eric.ed.gov/fulltext/ED542543.pdf

Freiberg, H. J., & Waxman, H. C. (1988). Alternative feedback approaches for improving student teachers' classroom instruction. *Journal of Teacher Education, 39*(4), 8-14. doi: 10.1177/002248718803900403

Gill, B., Shoji, M., Coen, T., & Place, K. (2016). The content, predictive power, and potential bias in five widely used teacher observation instruments. *Mathematica Policy Research.*

Retrieved from Regional Educational Laboratory website at http://ies.ed.gov/ncee/edlabs.

Government of Alberta. (2010). *Inspiring action on education.* Retrieved from https://www.oecd.org/site/eduilebanff/48763522.pdf

Hafen, C. A., Hamre, B. K., Allen, J. P., Bell, C. A., Gitomer, D. H., & Pianta, R. C. (2015). Teaching through interactions in secondary school classrooms revisiting the factor structure and practical application of the Classroom Assessment Scoring System–Secondary. *The Journal of Early Adolescence, 35*(5-6), 651-680. doi: 10.1177/0272431614537117

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development, 76*(5), 949-967. doi: 0009-3920/2005/7605-0001

Harms, T., & Clifford, R. (1998). *The Early Childhood Environment Rating Scale*. New York, NY: Teachers College Press.

Hattie, J. (2003). Teachers make a difference: What is the research evidence? *Distinguishing Expert Teachers from Novice and Experienced Teachers*. Camberwell: Australian Council for Educational Research. Retrieved from http://start.sd34.bc.ca/innovation/wp-content/uploads/2014/01/Hattie_TeachersMakeADifference.pdf

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Research Paper. MET Project.

Bill & Melinda Gates Foundation. Retrieved from http://files.eric.ed.gov/fulltext/ED540960.pdf

Klette, K., & Blikstad-Balas, M. (2018). Observation manuals as lenses to classroom teaching: Pitfalls and possibilities. *European Educational Research Journal, 17*(1), 129-146. doi: 10.1177/1474904117703228

La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The classroom assessment scoring system: Findings from the prekindergarten year. *The Elementary School Journal, 104*(5). http://dx.doi.org/10.1086/499760

McCombs, B. L. (2013). Educational psychology and educational transformation. In I. B. Weiner, W. M. Reynolds, & G. E. Miller (Eds.), *Handbook of Psychology, Volume 7, Educational Psychology* (2nd ed., pp. 493-533). New Jersey: John Wiley & Sons, Inc.

National Center for Teacher Effectiveness. (2012). *Mathematical quality of instruction.* Retrieved from http://isites.harvard.edu/icb/icb.do?keyword=mqi_training&tabgroupid=icb.tabgroup120173

O'Sullivan, M. K., & Dallas, K. B. (2017). A collaborative approach to implementing 21st century skills in a high school senior research class. *Education Libraries, 33*(1), 3-9.

Pakarinen, E., Kiuru, N., Lerkkanen, M. K., Poikkeus, A. M., Siekkinen, M., & Nurmi, J. E.

(2010). Classroom organization and teacher stress predict learning motivation in kindergarten children. *European Journal of Psychology of Education, 25*(3), 281-300. doi: 10.1007/s10212-010-0025-6

Patton, M. Q. (2011). *Developmental evaluation: Applying complexity concepts to enhance innovation and use.* New York: Guildford Press.

Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F. J. (2008). Classroom effects on children's achievement trajectories in elementary school. American *Educational Research Journal, 45*(2), 365-397. doi: 10.3102/0002831207308230

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109-119.

Pianta, R. C., Hamre, B. K., & Mintz, S. (2012). *Classroom Assessment Scoring System: Secondary manual.* Charlottesville, VA: Teachstone.

Pianta, R. C., La Paro, K., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS).* Baltimore, MD: Paul H. Brookes.

Puentadura, R. (2010). *Substitution, Adaptation, Modification, and Redefinition Model.* Retrieved from www.hippasus.com/rrpweblog/archives/2012/08/23/SAMR_BackgroundExemplars.pdf

Rose, D.H. & Meyer, A. (2006). *A Practical Reader in Universal Design for Learning.* Cambridge, MA: Harvard Education Press.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68. doi: 10.1037110003-066X.55.1.68

Sandstrom, H. (2012). The characteristics and quality of pre-school education in Spain. *International Journal of Early Years Education, 20*(2), 130-158. doi: 10.1080/09669760.2012.715012

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454-499. doi: 10.3102/0034654307310317

Smith, V., Hayes, S., Labonté, C., & Vargas, F. (2016). *Flexible pathways to success: Technology to design for diversity.* Retrieved from the Government of Alberta website: https://open.alberta.ca/dataset/029047d9-3a06-4b18-9c94-40d0c3937639/resource/e67821ad-ce9a-4f73-9248-b9710ebd6cf2/download/uofa-flexible-pathways-finalreport-july26-2016-online.pdf-

Stanford University. (2013). *PLATO: Protocol for Language Arts Teaching Observations.* Retrieved from

http://platorubric.stanford.edu/index.htmlhttp://platorubric.stanford.edu/index.html

Stuhlman, M. W., Hamre, B. K., Downer, J. T., & Pianta, R. C. (2010). Why should we use classroom observation? *Centre for Advanced of Teaching and Learning,* University of Virginia. Retrieved from https://curry.virginia.edu/uploads/resourceLibrary/CASTL_practioner_Part1_single.pdf

Vermette, D., Foote, D., Bird, C., Mesibov, D., Harris-Ewing, D., & Battaglia, D. (2001). Understanding constructivism(s): a primer for parents and school board members. *Education, 122*, 87–93.

Virtanen, T. E., Pakarinen, E., Lerkkanen, M. K., Poikkeus, A. M., Siekkinen, M., & Nurmi, J. E. (2018). A validation study of Classroom Assessment Scoring System–Secondary in the Finnish school context. *The Journal of Early Adolescence, 38*(6), 849-880. doi: 10.1177/0272431617699944

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research 73*(1), 89-122. doi:10.3102/00346543073001089

Westergård, E., Ertesvåg, S. K., & Rafaelsen, F. (2018). A preliminary validity of the classroom assessment scoring system in Norwegian lower-secondary schools. *Scandinavian Journal of Educational Research,* 1-19. doi: 10.1080/00313831.2017.1415964

Zohar, A., & Dori, Y. (2003). Higher order thinking skills and low achieving students: Are they

mutually exclusive? *Journal of the Learning Sciences, 12*, 145–181.

doi:10.1207/S15327809JLS1202_1

Chapter Two. The Classroom Assessment Scoring System – Secondary: A Systematic Scoping Review of an Observational Protocol of Classroom Practices

If "an education system is only as good as its teachers," it follows that teachers are vital to improving education (UNESCO, 2014, p. 3). Importantly, research continues to demonstrate teachers' classroom practices contribute significantly to student learning and can overcome known achievement gaps (e.g., Darling-Hammond & Rothman, 2015; Hamre & Pianta, 2005; Shyamalan, 2013). Beyond a student's pre-existing ability, Hattie (2003) proposed that teachers contribute 30% of the variance to academic achievement, suggesting quality teaching is the single most powerful external influence on students. However, delivering quality instruction is both a complex and challenging task (Downer, Jamil, Maier, & Pianta, 2012).

To better understand dimensions of quality instruction, methods to measure teachers' classroom practices have gained attention in educational research. As such, classroom observations have become increasingly more common (Wragg & Wragg, 2012), and a variety of measurement tools have been developed to target specific settings, subject areas (e.g., Harms & Clifford, 1998; National Center for Teacher Effectiveness, 2012; Grossman, Cohen, Ronfeldt, & Brown, 2014), or to capture global teaching practices in the classroom (e.g., Danielson, 2011; Pianta, La Paro, & Hamre, 2008). Although classroom observations have been conducted for

many years, it is only more recently that researchers have sought to develop valid and reliable standardized classroom observational protocols (Pianta & Hamre, 2009).

Of the various observational measures being used in research, one general pedagogical protocol, the Classroom Assessment Scoring System (CLASS: e.g., Pianta et al., 2008), has been widely adopted, particularly in the pre-school and elementary grades (Allen, Gregory, Mikami, Lun, Hamre, & Pianta, 2013). The CLASS has been used extensively in published research, for example, as an indicator of effective teaching, as a means to assess and provide feedback to teachers, or as a model to guide interventions. Based on developmental theory and supported by evidence of effective instructional practices or teacher-student interactional processes recognized to promote positive student outcomes (e.g., Battistich, Schaps, & Wilson, 2004; Emmer & Stough, 2001; Vermette et al., 2001; Zohar & Dori, 2003), the CLASS framework comprises three domains of effective teaching (Emotional Support, Classroom Organization, and Instructions Support; Pianta & Hamre, 2009). The CLASS was designed for trained observers to provide ratings of effective teacher classroom practices, while transcending boundaries related to content area. As such, there are versions that can be used in all types of classrooms from pre-kindergarten through high school (Pianta & Hamre, 2009). Each domain includes three to five dimensions and trained observers rate various behavioural indicators of teacher-student interactions for each dimension (Pianta & Hamre, 2009). One dimension, Student Engagement,

is not included in any of the domains. Importantly, although the three broad domains remain consistent across grade levels, the dimensions and behavioural indicators vary somewhat based on the developmental stage of the grade being observed (Pianta & Hamre, 2009).

The CLASS-S

Despite an explosion of research related to the CLASS in early and elementary years (e.g., a preliminary search returned 696 records), the measure adapted for use in middle and secondary classrooms was only published in 2012 (i.e., CLASS-Secondary [CLASS-S]; Pianta, Hamre, & Mintz, 2012). Grounded in theories of adolescent development and learning, the CLASS-S is specific to capturing teacher-student interactions in Grades 7 through 12 (Hafen, Hamre et al., 2015; Pianta et al., 2012). Following the same framework as the CLASS, the indicators of the CLASS-S place a greater emphasis on characteristics of adolescent development (e.g., adult and peer relationships, autonomy, competence; Deci & Ryan, 2000; Hafen, Hamre et al., 2015; Lavigne & Good, 2013). Nevertheless differences between the CLASS and CLASS-S have been described as minor, and one large scale project (the Measures of Effective Teaching project, Kane et al., 2012) used both the CLASS and CLASS-S to collect data from grade 4 though 9 classrooms and reported no disparities. Table 2.1 provides an overview of the domains and dimensions of the CLASS-S (column 'Version 5' displays the published form of the CLASS-S).

In order to become a certified observer, participants attend a two-day workshop and receive training from a master coder on how to rate each of the twelve CLASS-S dimensions using a 1-7 scale. Each dimension includes specific indicators, and observers are trained on how to identify and score each dimension. Scores in the 1-2 point range are classified as low, in the 3-5 point range as mid, and in the 6-7 point range as high (Pianta, Hamre, & Mintz, 2012). After training, observers must pass an online reliability test demonstrating they are able to code five video segments within one-point of a master coder, 80% of the time. According to the developers, a standard observation is based on watching a classroom lesson for approximately 15 minutes, followed by time to assign the ratings for each dimension, before beginning another observation (Pianta et al., 2012). The 15-minute observation period is also referred to as a "segment", and so each lesson may include multiple segments depending on the length of the class, if the lesson was video-recorded, or observed live.

The CLASS-S has been used across a breadth of published research reports, academic articles (i.e., peer-reviewed), unpublished dissertations, and grey literature. As the body of research using the CLASS-S grows, the author of the current systematic scoping review proposes now is the time to pause and critically examine the evidence base for its use and make recommendations for future research. The application of the CLASS-S as a standardized, valid and reliable observational tool deserves attention in the current era where educational policy

makers and administrators seek methods to be accountable and to raise standards in schools through professional development. In addition, the educational community has acknowledged the need to identify and support effective teacher classroom practices to enhance student outcomes in an attempt to close the achievement gap across North America and beyond.

What is a Scoping Review?

The purpose of a scoping review is to define the parameters of a given topic by systematically taking stock of available research (Grant & Booth, 2009; Peters, Godfrey, Khalil, McInerney, Parker, & Soares, 2015). The resultant "mapping" of research includes capturing the full range of evidence (including quantitative and qualitative data) available pertaining to a specific topic (Peters et al., 2015). The goals of a scoping review are to define key themes, summarize and disseminate research findings, and to make recommendations based on identified gaps in knowledge (Arksey & O'Malley, 2005; Peters et al., 2015). Similar to a systematic review, the methods of a scoping review are rigorous, transparent, and replicable to increase confidence in the findings (Arksey & O'Malley, 2005; Grant & Booth, 2009).

A systematic scoping review differs from a systematic review in that the latter is driven by highly specific research question(s), and typically includes the application of quality ratings for each of the included studies. In contrast, a systematic scoping review provides an opportunity to summarize the existing evidence associated with a more general research

question. This methodology is particularly helpful for the current topic, as it allows for the systematic summarization and integration of emerging evidence when there is not a larger body of research addressing a similar topic or outcome.

Study Purpose

As described previously, the CLASS-S is an observation tool of classroom practices that is being used widely in research and practice. The purpose of the present study is to review the current body of emerging literature using the CLASS-S in order to summarize the key themes, research findings, and to make recommendations. Specifically the aim of the present systematic scoping review is to:

1. Provide an overview of the CLASS-S in published literature (e.g., who is using it, how are they using it?).
2. Map key themes addressed by authors using the CLASS-S.
3. Present findings describing relationships between classroom (i.e., CLASS-S) and student outcomes (e.g., student achievement).
4. Identify gaps in the literature and summarize recommendations for the future use of the CLASS-S.

Methodology

Methodology for conducting a systematic scoping review as outlined by Arksey and

O'Malley (2005) guided the present paper. Specifically, a five stage methodological framework was followed: 1) establish the research question(s), 2) search relevant studies, 3) select studies for inclusion, 4) extract and chart the data, and 5) synthesize and report results. Furthermore, various methods and procedures described by Peters and colleagues (2015) informed the present paper. For example, recommendations for how search procedures were reported, and suggestions for inclusion of items in the data extraction form (e.g., source origin/country of origin, aims/purpose, methodology) were adopted.

Method

Search Strategy

Two independent reviewers conducted comprehensive searches of published literature (including peer reviewed and grey literature). Appropriate electronic databases for these searches were identified in consultation with a librarian with subject expertise in the field of education research. In September 2017, electronic databases including Education Research Complete, Education Research Abstracts, ERIC (via EBSCOhost), Academic Search Complete, PsycINFO, Web of Science, Scopus, and ProQuest Dissertations and Theses Global (see Figure 2.1) were searched. Studies published in any language were eligible. One identified study published in Dutch was not available for review (Wubbels, 2014) despite requesting an inter-library search, and emailing the author. The general search terms used were (*Classroom*

Assessment Scoring System) AND (*Secondary*). Furthermore, other relevant sources (e.g., Gates Foundation and Teachstone websites) and reference lists from studies identified for inclusion were hand searched.

Inclusion Criteria

After duplicates were removed, two independent reviewers screened remaining publications, and 101 articles were identified for further assessment of eligibility. If only one reviewer identified an article through screening for further assessment, it was included for subsequent review. Once articles were read thoroughly, there were no disputes regarding final inclusion in the scoping review. Figure 2.1 provides a flow diagram of Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) according to methodological guidelines (Moher, Liberati, Tetzlaff, Altman, & the PRISMA group, 2009; Peters et al., 2015). Criteria for inclusion were: (1) use of the CLASS-S to observe students; (2) classroom observations of students from middle and/or high school (i.e., Grades 7-12); and (3) numeric CLASS-S scores of observations reported. Criteria for exclusion were: (1) no use of the CLASS-S; (2) classroom observations did not include any students in Grade 7 or higher; (3) not enough demographic data was provided to assess the inclusion criteria; (4) study noted it drew from the constructs of the CLASS-S, but did not actually use the measure; (5) review papers or book

chapters without original research; or (6) advanced publication, research brief, or magazine article based on original research already included in the review.

Data Extraction and Charting

An a priori data extraction form was created based on recommendations for systematic scoping reviews (Peters et al., 2015) and minor revisions were made in an iterative process following the review of the first few publications. In particular, it was identified that articles reported on slightly different versions of the CLASS-S (e.g., pilot, unpublished, or published) and so these were subsequently tracked on the data extraction form. The final revised data extraction form was provided to the second reviewer, who was asked, but did not identify a need for any additional revisions. Final data extracted included: location of the study, authors affiliation with CLASS-S developers, CLASS-S version used, research design and methods, study relationship with other included data/articles, participating teacher and student demographics (sample size, grades, teaching experience), CLASS-S coding information (coder demographics, methods, reliability, descriptive data), study purpose, and overall findings specific to the CLASS-S. As this review was a scoping review and sought to capture the breadth of use of the CLASS-S in published literature, no indicator of methodological quality was calculated (Arksey & O'Malley, 2005; Peters et al., 2015).

One reviewer independently completed the data extraction form for each of the included

studies. The second reviewer completed data extraction for the first five included articles and met with the first reviewer to clarify any process questions and build consensus. As no further changes were required to the data extraction form, and consensus was reached, the second reviewer continued to apply the form to the remaining articles. Once all data had been extracted, the two reviewers met to compare each cell of the data extraction form, any inter-reviewer discrepancies were discussed and resolved by consensus. Following discussions of any discrepancies, the reviewers were in 100% agreement of the completed data (included articles and selected characteristics are presented in Table 2.2).

Results

This section provides search results and an overall "mapping" of the research. Results include an overview of the CLASS-S in published research, a summary of key themes, and research findings specific to student outcomes.

1. Overview of the CLASS-S in Published Research

Fifty-five publications were identified for inclusion (see Figure 2.1). Figure 2.2 presents the total number of studies by year of publication. The first published study using the CLASS-S appeared in 2009. This figure does not include the two articles published in 2017, as the literature search was completed in early September 2017 and, therefore, would not accurately

reflect the total number of publications that may follow in the remainder of 2017. In the event that an article was published early online, the final or print publication date was used.

Figure 2.2 also includes a breakdown of the total number of publications authored by, or associated with, one of the developers of the CLASS-S. Association was identified as an author who had co-authored a publication(s) with one or more of the original CLASS-S developers. Approximately half (51%) of all identified studies were authored by at least one individual associated with the original development of the CLASS-S. However, Figure 2.2 illustrates that since 2015, publications written by authors not associated with the CLASS-S developers have begun to outnumber those published by associated authors. For example, authors not associated with the developers wrote 40% of the publications in 2014, 50% of the publications in 2015, and 71% of the publications in 2016. Furthermore, of the two publications included in this study from 2017, none of the authors were known to be directly associated with the original CLASS-S developers.

Overall, the identified publications included 38 journal articles (69%), six book chapters (11%), five research reports (9%), four doctoral dissertations (7%), and two conference proceedings or papers (4%). The 38 articles were published in 30 unique journals, with most journals publishing only one article (83%), while three journals published two papers (10%), and

two journals published more than two papers (7%). Of the book chapters, five (83%) were published in one book.

The CLASS-S has been used in publications originating from Australia, Finland, Norway, United Kingdom, and the United States (see Figure 2.3). However, the majority (87%) of identified studies were conducted in the United States. Figure 2.3 provides more details about the location (i.e., state) in the United States where data was collected. In some instances, authors did not specify exactly where data was collected, for example they may have provided a broad location like in a "northwest state," this information was represented by the "unspecified" location category in the figure. When publications included data from a large-scale project, such as the Measures of Effective Teaching (MET) project with data collection from various states, all states known to have been a location for the project were identified, unless authors noted drawing from a specific subsample of data.

Publications used different versions of the CLASS-S, as demonstrated in Table 2.1. As is common with a tool under development, some changes occurred with the domains and dimensions of the protocol. The largest change saw the Negative Climate dimension move from the Emotional Support domain in earlier iterations, to the Classroom Organization domain in the most recent two versions. The move was based on factor analysis and resulted in the final version of the CLASS-S that was subsequently published (see Hafen, Hamre, et al., 2015). Of

note, only 10 articles have used the CLASS-S version where Negative Climate was included in the Classroom Organization domain. This is noteworthy because the Negative Climate dimension tends to reflect less variability (e.g., less than 1% scored in the ineffective range; Kane & Staiger, 2012). Of these 10 articles, two used the final published version of the CLASS-S which also included a name change of the Analysis and Problem Solving dimension to Analysis and Inquiry, and added the Instructional Dialogue dimension. Therefore, the final published version of the CLASS-S has been used in less than 4% of all published research.

Many publications included data from larger projects. A larger project was identified (see Table 2.2) when two or more studies drew from the same data set. Of note, some articles drew data from more than one larger project. Total articles based on data from identified larger projects included: Ten (18%) from the MET project, seven (13%) from My Teaching Partner – Secondary, five (9%) from My Teaching Partner – Secondary replication study, six (11%) from Toward an Understanding of Classroom Context, four (7%) from Understanding Teaching Quality, three (5%) from The National Center on Scaling up Effective Schools, three (5%) from Urban College Academy, two (3.5%) from Assessing Induction and Mentoring and, two (3.5%) from eMINTS. Furthermore, 18 articles were published using independent data sets. Therefore, a total of 27 unique data sets informed the publications in the present review.

Coding procedures varied across articles, with many researchers reporting on video

recorded classroom sessions (30 articles; 54.5%), live classroom sessions (14 articles; 25.5%), or a combination of video recorded and live sessions (seven articles; 13%). Coder training was reported in 41 articles (74.5%), with the majority of researchers indicating coders attended a CLASS-S training workshop and completed the publishers online reliability testing. Two articles (Malmberg, Hagger, Burn, Mutton, & Colls, 2010; Virtanen, Lerkkanen, Poikkeus, & Kuorelahti, 2015) included training in live classrooms; however, only Virtanen et al., 2015 used live classroom scoring in their research procedures. Therefore, some research included live classroom scoring, although coders were not trained using experiences in live classroom scoring.

Nineteen articles (34.5%) provided some demographic information about the coders. For example, coders included advanced undergraduate or graduate students, former secondary school teachers, supervisors in a pre-service teacher education program, or program research coordinators. The number of coders per classroom segment varied across studies and was difficult to capture accurately as some articles reported on more than one project with different coding procedures. Double coding classroom segments appeared to be the preference of the majority researchers, and 13 articles (24%) noted double coding of all observations, while three articles (5%) reported coding all segments only once. The remaining studies indicated a sample of classroom segments were double coded, with the proportion ranging from 5% to 75%.

Despite researchers typically coding classroom segments more than once, reliability procedures used to minimize coding drift or rater error were not commonly reported. Procedures to check for or maintain reliability (i.e., calibration exercises, meetings to jointly code master tapes) were reported in 15 articles (27%). In contrast, inter-rater reliability statistics were reported in 31 articles (56%). Examples of statistics included intra-class correlation coefficients (14 articles; 25.5%), percentage of agreement within one point (five articles; 10%), Cronbach's alpha (five articles; 10%), and Cohen's kappa coefficient (three articles; 5.5%). Although procedures were not always put in place to minimize coding drift, reliability statistics were more often provided and generally fell within acceptable ranges.

2. Mapping Key Themes

Table 2.2 identifies key themes and provides a brief description of findings specific to the CLASS-S for each article included in this review. Findings across articles fell in eight broad thematic categories (see Figure 2.3 for a graphic representation). Key themes included: 1) relationships between classroom and student outcomes (e.g., student achievement, engagement, and perceptions of classroom context), 2) articles specifically addressing the CLASS-S model (e.g., factor structure, concurrent validity, or measurement error), 3) school/classroom contrasts (e.g., differences between schools identified as high value-added vs. low value-added, honours track vs. regular track classrooms, middle school vs. high school, or a public vs. a charter

school); 4) CLASS-S procedures and the process of coding (e.g., number and length of observation segments, training of coders, or calibration exercises), 5) teacher/curricular interventions (e.g., CLASS-S informed intervention impact on teaching practices, pre- to post-intervention CLASS-S scores), 6) teacher characteristics (e.g., CLASS-S scores improved with experience in their first two years of professional practice, or teacher emotional intelligence was associated with higher scores on the Regard for Adolescents domain), 7) case studies (e.g., CLASS-S scoring observation notes coded qualitatively, or in a mixed methods approach), and 8) teacher education program assessments (e.g., linking CLASS-S scores to pre-service teaching self-efficacy beliefs, or ratings of extraversion and depression).

In summary, 20% of identified publications presented findings related to relationships between classroom (i.e., CLASS-S) and student outcomes. A further 20% of articles addressed the CLASS-S model. School/classroom contrasts were described in 13% of articles, with an additional 13% of publications presenting findings specific to CLASS-S procedures and the process of coding. Teacher/curricular interventions were assessed in 14% of included articles, while teacher characteristics were addressed in 10% of the sample articles. Researchers used the CLASS-S in 6% of articles using a case study format, and 4% of articles provided findings specific to teacher education program assessments programs.

3. Findings Describing Relationships Between Classroom (i.e., CLASS-S) and Student Outcomes

Articles were identified by the key theme "relationships between classroom and student outcomes" (see Table 2.2) when authors reported on findings linking the CLASS-S to student outcomes. Overall, authors of 15 articles described relationships between classroom and student outcomes, representing 27% of all CLASS-S publications summarized in this review. Specifically, student outcome themes addressed student achievement, student engagement, and student perspectives of their classroom context.

Student achievement. The most common student characteristic addressed by researchers was the association between the CLASS-S and student achievement. Student achievement was measured in a variety of ways, by scores on state-mandated standardized subject tests (e.g., Allen et al., 2013; Allen, Hafen, Gregory, Mikami, & Pianta, 2015; Allen, Pianta, Gregory, Mikami, & Lun, 2011), a unit test (Culp, Martin, Clements, & Presser, 2015), by current grade point average (GPA; e.g., Yoder, 2013), gains from a pre- and end of course algebra test (e.g., Bell et al., 2012), or by value-added models incorporating conceptual mathematics and open-ended literacy assessments (e.g., Kane & Staiger, 2012; Mihaly & McCaffrey, 2014; Ruzek, Hafen, Hamre, & Pianta, 2014). Value-added models attempt to statistically quantify the impact a teacher has on student achievement, after controlling for various factors such as student outcomes (e.g., past

achievement scores) and school/classroom characteristics (e.g., class size, peer factors; Koedel, Mihaly, & Rockoff, 2015). The following is a summary of findings linking the CLASS-S and student achievement and summative measurement, intervention studies, and value-added models.

Student achievement and summative measurement. Only one article identified (Allen et al., 2013) was designed specifically to analyze the relationship between the CLASS-S and student achievement. Drawing on data from 37 classrooms (643 students) across six school districts, Allen and colleagues (2013) employed hierarchical linear modeling (HLM) to examine the relationship between the CLASS-S and student achievement. The researchers measured student achievement by state-mandated standardized subject test scores, while controlling for other contextual factors such as student grade, gender, socioeconomic status, and classroom size, known to impact student achievement. Researchers found that all three CLASS-S domains predicted end of year student achievement, after controlling for prior year achievement scores, and other student and classroom characteristics. Of the CLASS-S domains, Emotional Support was identified as having the strongest predictive value on increased student achievement. The authors described that if a student with prior test scores in the average range (i.e., 50th percentile) was enrolled in a classroom where Emotional Support was rated higher (i.e., one standard deviation above the mean), the student's end-of-year achievement score was predicted to fall at the 59th percentile. In contrast, if the same student were enrolled in a classroom where

Emotional Support was rated lower (i.e., one standard deviation below the mean), their end-of-year achievement score was predicted to fall at the 41st percentile.

When explored further, researchers identified scores on the Positive Climate, Teacher Sensitivity, Regard for Adolescent Perspectives, Instructional Learning Formats, and Analysis and Problem-Solving dimensions predicted the greatest student achievement. In addition, researchers identified prior student achievement scores predicted higher scores on the Behavior Management, Instructional Learning Formats, Content Understanding, and Quality of Feedback dimensions. Therefore, findings suggested pre-existing student outcomes (i.e., prior student achievement) predicted higher scores on dimensions of the Classroom Organization and Instructional Support domains, but not the Emotional Support domain. The authors concluded classrooms with higher levels of Emotional Support led to greater year-end gains in student achievement, regardless of students' prior academic achievement scores.

Furthermore, the authors analyzed the data for interactions that moderated associations between the CLASS-S and achievement scores, and identified class size interacted with both the Emotional Support and Instructional Support domains. Emotional Support and Instructional Support domain scores predicted greater end-of-year achievement in smaller classrooms (i.e., approximately 17 students), in contrast to larger classrooms (i.e., approximately 29 students).

In another article that utilized a summative indicator of student achievement, Yoder (2013) examined associations between the CLASS-S and self-reported GPA. Yoder (2013) asked students to indicate their current grade in the subject where he also conducted observations using the CLASS-S. No associations between CLASS-S and student achievement were identified in this study, and Yoder (2013) suggested his results were limited due to a small sample size (i.e., 19 classrooms). However, methodological challenges, including the use of self-reported GPA, and that the author was also the only CLASS-S coder, may have further impacted the outcomes reported.

Student achievement and intervention studies. Three articles explored the impact of intervention by examining the relationship between changes in CLASS-S scores and student achievement. Specifically, two randomized trials measured the impacts of My Teaching Partner – Secondary (MTP-S; Allen et al., 2011; Allen et al., 2015), and one study explored the impact of the Exploring Photosynthesis program (Culp, Martin, Clements, & Presser, 2015).

Of the two MTP-S studies; one included mostly middle school teachers (78 classrooms; Allen at al., 2011), while the replication study included mostly secondary school teachers (86 classrooms; Allen et al., 2015). MTP-S provided middle and secondary school teachers with individualized, web-mediated coaching, to support their classroom practices (Allen et al., 2011, Allen et al., 2015). Researchers in both studies concluded teacher participation in MTP-S

resulted in student achievement gains across subject areas, when compared to teachers who did not participate in the intervention. Gains were reported by the researchers to translate to an increase on achievement tests from the 50th percentile to the 59th percentile for the average student.

In the other article that addressed the relationship between the CLASS-S and student achievement in an intervention, researchers explored the impact of the Exploring Photosynthesis program, a technology-based intervention (i.e., digital games; Culp et al., 2015). Forty-one classrooms were randomly assigned to either the intervention (21 classrooms) or control group (20 classrooms), and all classrooms were observed using the CLASS-S protocol, and student achievement was assessed by a unit test. The authors reported an interaction effect between CLASS-S scores and student outcomes for the classrooms in the intervention group. Researchers indicated teachers who scored higher on the CLASS-S and used the Exploring Photosynthesis program had better student achievement on the unit test. The authors hypothesized more effective teachers were able to implement technology to enhance learning by making meaningful connections for students.

Student achievement and value-added models. Three additional articles based on data collected from over 3000 classrooms for the MET project (Kane & Staiger, 2012; Mihaly & McCaffrey, 2014; Ruzek et al., 2014), addressed the relationship between the CLASS (including

both the CLASS and CLASS-S) and student achievement. In contrast to studies that used a single, summative measure of student achievement, these articles measured student achievement gains using value-added scores, an estimate of a teacher's contribution to student achievement gains that controls for other factors (see Kane & Staiger, 2012, and Mihaly, McCaffrey, Staiger, & Lockwood, 2013, for model details). Researchers developed value-added scores as a means to statistically capture the effects of teaching. For example, in the MET project, "[a] teacher's average student achievement gain is the average difference between students' actual and expected achievement test score at the end of the year across all tested students in a classroom who have a prior year achievement test score" (Kane & Staiger, 2012, p. 40). The MET project included scores from state mandated subject tests, as well as supplemental assessments of mathematical concepts and literacy.

Overall findings of the MET project were summarized in a research report authored by Kane and Staiger (2012). In brief, researchers with the MET project found higher CLASS scores were moderately associated with gains in student achievement (as measured by gains on state standardized assessments). The researchers identified teacher's scores varied greatly from lesson to lesson, and therefore recommended ratings on observational protocols be averaged across multiple raters, and multiple lessons in order to promote reliability. In addition, observation scores became a stronger predictor of student gains when student achievement scores were

combined with student survey data, and even stronger when combined further with value-added scores (e.g., gains on standardized mathematics and literacy assessments). Furthermore, teachers with higher combined scores had students that reported more classroom effort and enjoyment.

In an article examining CLASS data from the first year of the MET project, researchers Ruzek and colleagues (2014) used the overall CLASS score as an indicator of instructional quality, and the value-added score as an indicator of teacher's impact on student achievement, to explore different strategies assessing the relationship between these two indicators. Three of the four strategies presented in the article assumed a linear relationship between the CLASS and value-added scores. However, the fourth strategy used statistical analysis (i.e., spline regression) to identify thresholds, or breakpoints, where associations between the CLASS and value-added scores differed. Findings from the strategy using spline regression indicated the association between CLASS observation scores and student achievement was nonlinear (particularly if the CLASS score fell in the mid range, 2.91-4.5). Furthermore, the researchers identified "active ranges" (above 4.5, and below 2.91) where CLASS scores were associated with greater gains in value-added scores. The authors posited targeting professional development for teachers in the bottom range, or those closest to the active range (i.e., teacher with a CLASS score of 4), may provide the greatest gains in value-added scores.

In another article based on a subsample of data from the MET project, Mihaly and

McCaffrey (2014) examined grade-level differences in three of the six districts included in the larger data set (i.e., New York City; Charlotte, North Carolina; and Hillsborough, Florida). The authors reported teachers in Grades 4 and 5 scored higher than Grade 6 teachers (elementary ratings were based on the CLASS-Upper Elementary version), and Grade 6 teachers scored higher than teachers in Grades 7 and 8 (middle school ratings were based on the CLASS-S). In further analysis of the Hillsborough, Florida data (as it was considered the most complete), the authors' explored grade-level differences and student achievement based on two different teachers' value-added scores for both state-mandated subject tests, and alternative assessments (i.e., conceptual mathematics and open-ended literacy assessments). Despite higher CLASS scores in elementary grades, the authors reported the CLASS was not a stronger predictor of student achievement in elementary grades, as compared to middle school grades. Meaning that although CLASS scores were higher in elementary grades, associations between these scores and student achievement were not stronger. Furthermore, the authors suggested lower scores in middle schools might be related to yet undetermined factors (i.e., student outcomes, teacher effectiveness).

In contrast, authors of an article using value-added models based on an independent dataset, Bell and colleagues (2012), reported divergent evidence linking the CLASS-S to student achievement. The researchers measured student achievement using a value-added model derived

from pre- and end of course algebra test scores across 82 classrooms. Specifically, they found the domains of Classroom Organization and Instructional Support were associated with value-added scores; however, when they adjusted the model in an effort to separate teachers' contributions from student contributions, only Classroom Organization remained statistically significant. Furthermore, the authors identified greater gains when the value-added model accounted for pretest scores, suggesting that greater gains were made when students were better prepared. The authors propose the CLASS-S detects classroom quality that incorporates both teacher and student factors. The authors acknowledged their findings were different from other published results, and noted limitations with sample size.

Student engagement. The second most common student characteristic addressed by researchers was the association between the CLASS-S and student engagement. Student engagement was measured in a variety of ways, by student-ratings (Hafen et al., 2012; Ruzek et al., 2016; Virtanen, Lerkkanen, Poikkeus, & Kuorelahti, 2015; Yoder, 2013), by teacher-ratings (Virtanen, Lerkkanen, Poikkeus, & Kuorelahti, 2015), and by observation scores on the CLASS-S (Hafen et al., 2012; Malmberg et al., 2010; Virtanen, Lerkkanen, Poikkeus, & Kuorelahti, 2015). In a study examining eight classrooms, Virtanen and colleagues (2015) compared all three ways of rating student engagement and found that all measures (i.e., student-ratings, teacher-ratings, and observations) were associated with higher scores across all domains of the

CLASS-S. Of the domains, Classroom Organization was associated with all three measures of engagement, while Instructional Support was associated with student-ratings and observations (but not teacher ratings). Emotional Support had an indirect effect on student engagement via Classroom Organization and Instructional Support.

Researchers in three other articles used student-ratings of engagement. Hafen and colleagues (2012) reported student-rated and observed CLASS-S Student Engagement were predicted by classroom-level perceptions of autonomy at the beginning of the year in a sample of 34 classrooms. Although student-ratings and CLASS-S observed engagement was not correlated at the beginning of the year (time 1), they were strongly associated by the end of the year (time 2). The authors reported observed engagement and student self-reports became more comparable over the course of the school year. In another article exploring engagement in 68 classrooms throughout the year, Ruzek and colleagues (2016) indicated classrooms with higher Emotional Support scores on the CLASS-S in the fall had students who reported increased behavioral engagement and mastery motivation at the end of the school year. Students also reported more autonomy and positive relationships with their peers at the mid-point of the year when they were in classrooms with higher fall observed Emotional Support. The researchers identified an indirect effect of Emotional Support on engagement and motivation, mediated by autonomy and peer relatedness. In the final article included in this review to use student-ratings

of engagement, Yoder (2013) reported no association between self-reported engagement and the CLASS-S.

In the only included article to discuss student engagement solely based on CLASS-S observations, Malmberg and colleagues (2010) used multi-level modeling to examine data at the segment, time point, and teacher levels in 17 classrooms. The authors identified that at the segment level student engagement strongly predicted all three domains. However, when the researchers explored variability in scores across lessons, they found student engagement varied from lesson to lesson, but was no longer associated with Emotional Support Furthermore, variability in student engagement was not related to variability in Instructional Support. Authors also found larger classrooms (class sizes included in the study ranged from four to 35 students) were associated with increased student engagement. This finding seems to conflict with past research suggesting smaller classes are linked to greater student engagement (Ehrenberg, Brewer, Gamoran, & Willms, 2001). Yet these findings may reflect underlying theories of adolescent development, particularly the importance of peer relatedness (Hafen et al., 2012), of which the CLASS-S is based upon.

Student perspectives of classroom context. Researchers sought to link the CLASS-S with students' perspectives of classroom contextual factors (i.e., classroom structure, classroom climate). In one article, Spearman and Watt (2013) examined classroom context and its impact

on girls' motivation for science across a sample of five classrooms. The authors asked students to report on their perceptions of teaching style, motivation for science, and conducted classroom observations using the CLASS-S. They reported Positive Climate, Negative Climate, Regard for Adolescent Perspectives, Behavior Management, Quality of Feedback, and Student Engagement all impacted a student's perception of structure in the classroom, which in turn indirectly impacted a student's reported extrinsic utility value for science. Researchers also compared students' responses on a questionnaire of teaching styles that closely paralleled the six CLASS-S dimensions included in the study. They found girls who rated a dimension of teaching style as low and attended a class where the parallel dimension received low CLASS-S ratings had lower levels of motivation for science. However, girls who rated a dimension as high and were observed in a classroom where the dimension was also rated as high on the CLASS-S, motivation for science decreased from the start of the school year. The researchers suggested findings might reflect a ceiling effect. They also reported girls who rated teaching style dimensions highly at the start of the year, yet were in classrooms where the parallel CLASS-S dimension was rated lower, reported the highest levels of motivation at the end of the year.

In the second article to examine student perspectives and classroom context, Yoder (2013) completed classroom observations using the CLASS-S and asked students to report on classroom climate. Classroom climate was assessed based on measures of academic and

emotional peer support, teacher-student relationship, and classroom misbehavior. The author reported classroom observation scores on the CLASS-S were not related to any measured variables of classroom climate and cited possible limitations due to sample size. However, there were likely also limitations related to the author's method of measuring classroom climate.

Discussion

Identify Gaps and Summarize Recommendations

This scoping review presents an overview of the breadth of studies using the CLASS-S as a standardized observational protocol of classroom practices, maps key themes addressed across studies, and summarizes study findings specific to student outcomes. Search results identified 55 publications since 2009 that met inclusion criteria. Examining the use of the CLASS-S in publications identified gaps in knowledge about its psychometric properties, its association with key themes (e.g., student achievement), and its ability to enhance classroom practices. Importantly, the CLASS-S has emerged as a potential tool to deepen our understanding of the impact of classroom practices in middle and high school settings. However, more research is needed to support its use.

Gap 1: Psychometric properties. Although in general, publications using the CLASS-S have been declining (i.e., only two studies published in 2017 qualified for inclusion in this review), trends suggest the broader research community have begun to use the CLASS-S (see

Figure 2.1). However, less than 4% of studies included in this review used the final version of the CLASS-S published in 2012. Therefore, the majority of published research used earlier iterations of the CLASS-S. Furthermore, many publications have collected data in the United States, with fewer articles including international data.

As described in the results section of this review, the majority of articles reported standard certification training procedures for CLASS-S coders. However, some studies did not provide training in live classrooms prior to completing live coding. Similarly, many authors reported double coding all, or a proportion of classroom segments, yet there appears to be no consensus on best practice regarding double coding. Some researchers (Burchinal, 2018) have critiqued the fact that certification standards state coders are required to score within one point of the trainer 80% of the time. Furthermore, Burchinal (2018) noted typical standard deviation for each item is one point or less, meaning a range of one standard deviation, 80% of the time, is considered acceptable. As a result, a great deal of variance is likely attributable to differences between coders.

Despite recommendations outlined by the American Educational Research Association (AERA) for the use of observation protocols in the classroom (AERA, 2015), it is not clear if the CLASS-S meets all the reliability and validity requirements as defined in the Standards for Educational and Psychological Testing (AERA, American Psychological Association [APA], &

National Council on Measurement in Education [NCME], 2014). Typically, authors of articles included in this review have examined reliability by assessing inter-rater reliability, or the degree of agreement between coders. Although methodological quality was not examined in this review, Hallgren (2012) has identified general concerns with how studies typically assess inter-rater reliability in observational data. Hallgren (2012) stated the most common errors include the use of incorrect statistical analyses, misinterpretation, and neglecting to consider the impact of inter-rater reliability on subsequent analyses. One common example often used in studies included in this review, was the inclusion of percentages of agreement as an indicator of IRR. According to Hallgren (2012), percentage of agreement has not been supported for use as a measure of IRR. One criticism noted is that this method typically does not correct for agreements that would be expected by chance.

In regard to validity, it is based on an accumulation of supporting evidence for every interpretation of a given test or measure (AERA, APA, & NCME, 2014). As identified in this review, the CLASS-S can be used in a wide variety of ways. For example, the CLASS-S might describe a teacher's current level of teaching effectiveness, provide an outcome indicator for interventions, or contribute to future predictions about student achievement in value-added models. According to the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014), evidence from a variety of categories is needed to support each use. Relevant

categories of evidence to support the validity of the CLASS-S include validity of structure, content, relationships with conceptually related constructs, and relationships with criteria.

Approximately 20% of published articles have reported on the CLASS-S model, and some researchers have provided emerging evidence to support its validity (i.e., structural validity; e.g., Allen et al., 2013; Hafen, Hamre, et al., 2015), and reliability (e.g., Kane et al., 2012). Conversely, other researchers have challenged the three-factor structure in favor of a single-factor as the best indicator of teaching effectiveness based on the CLASS-S (Kane et al., 2012; Malmberg, et al., 2010; McCaffrey et al., 2015). In the literature, the developers reported content-oriented evidence (see Pianta, Hamre, & Mintz, 2012); however, researchers exploring the opinions of novice teachers of the relevance of the CLASS-S constructs and if they had learned them in pre-service teacher education programs, were unable to provide corroboration (Caspersen & Raaen, 2014). Further, some researchers have questioned CLASS-S content, for example, Gamlem and Munthe (2014) found that the scoring criteria for Quality of Feedback emphasized encouragement-oriented over learning-oriented feedback. Furthermore, limited evidence describing relationships between the CLASS-S and other related constructs have been identified (e.g., Frameworks for Teaching [Gill et al., 2016], Teacher Emotional Intelligence Measure [Friedman, 2014]). Finally, evidence is needed to support the relationship between CLASS-S scores and predictions of performance, or the relationship with criteria. Specifically,

is there evidence to support the CLASS-S ability to discriminate and predict based on criteria associated with its use (e.g., as an indicator of effective teaching)? The relationship between the CLASS-S and key constructs (e.g., student achievement) will be discussed in the section, *Associations with Key Themes*. Results of this review suggest evidence is emerging in these various categories; however, there remains a need to build on existing evidence to support each of the potential uses of the CLASS-S.

Recommendation 1: Further research to support psychometric properties. Despite anecdotal reports of an increase in the CLASS-S in international research and practice (L. Sernett, personal communication, February 8, 2018), there remains a need for published articles addressing its psychometric properties and providing validity evidence for each use. In particular, more research is required using the final published version of the CLASS-S. Establishing the CLASS-S as a protocol for observing classroom practices across international boundaries would strengthen our understanding of fundamental classroom instructional practices that transcend content and geographic borders. It should be noted that since the search for article inclusion for this review was conducted, two articles examining the validity of the CLASS-S in international contexts have been published (i.e., one from Finland [Virtanen1, Pakarinen, Lerkkanen, Poikkeus, Siekkinen, & Nurmi, 2017] and from Norway [Westergård, Ertesvåg & Rafaelsen, 2018]). In addition, addressing questions of best practice for double coding and

discussing common errors in inter-rater reliability procedures and the CLASS-S remains requires more attention. Moreover, independent research supporting the validity of content, and relationships with related constructs and other criteria is needed to support each of the common (and other potential) uses of the CLASS-S (e.g., as a measure of classroom quality, as a predictor of student achievement, as an indicator of effective teaching, and as a model and measure for teacher education programs).

Finally, researchers suggested different methods are needed to explore the complexity of the relationships among dimension scores, in contrast to typical rank or item level findings (e.g., Halpin & Kieffer, 2015). For example, Halpin and Kieffer described an approach to latent class analysis, and identified four distinct profiles of instructional practice demonstrating statistical relationships among items. The authors suggested latent class analysis might be a potential method for future research. Another example of exploring CLASS-S data was described by Ruzek, Hafen, Hamre, and Pianta, (2014), where they used spline regression to identify active ranges to categorize teachers. Given the complexity of classroom observations, new methods exploring the psychometrics of the CLASS-S and ways to apply it to classroom observation data are recommended to strengthen its use in research.

Gap 2: Associations with key themes. Mapping key themes across articles in this review identified eight categories (see Figure 2.4). Across the categories, although some

findings seemed to be consistent despite limited research (i.e., two articles identifying similar findings), divergent findings were revealed within the categories. For example, researchers examining school/classroom contrasts identified higher overall CLASS-S domain scores (Donaldson, LeChasseur, & Mayer, 2017; Smith, Preston, Haynes, & Neergaard Booker, 2015), or Emotional Support scores (Rutledge, Cohen-Vogel, & Osborne-Lampkin, 2012) in high track (or honours track) classrooms when compared to regular track classrooms. Similarly, other researchers reported higher CLASS scores in elementary classrooms when compared to middle school classrooms (Mihaly &McCaffrey, 2014), and the same pattern was reported for the Emotional Support domain only (Shell, Gazelle, & Faldowski, 2014). Yet other researchers found contrasting results, for example authors have identified increased Content Understanding and Analysis and Problem Solving scores in mathematics classrooms as compared to English language classrooms (Donaldson, LeChasseur, & Mayer, 2017), while others identified mathematics classrooms scored lower overall on the CLASS-S than English language classrooms (Mihaly & McCaffrey, 2014). Therefore, some patterns appear to be enduring, such as differences in high versus low track classrooms, and between elementary and middle school grades; however other patterns are not clear. In addition, some authors note no major differences on the CLASS-S when comparing Low Value-Added (LVA) to High Value-Added (HVA) schools (Grossman, Loeb, Cohen, & Wyckoff, 2013; Rutledge et al., 2012, Smith et al., 2015),

while others found higher Emotional Support (not including the Positive Climate dimension) in HVA schools (Smith, Cannata, & Haynes, 2016). Importantly, this pattern contrasts with other findings where researchers identified higher individual teacher scores on the CLASS-S were associated with greater value-added scores (Kane & Staiger, 2012).

Similar to school/classroom contrasts, other mapped key themes, such as teacher characteristics, provided some emerging, and some inconsistent findings. For example, the pattern of scores and growth across CLASS-S domains in novice teachers found CLASS-S scores improved in the early years of professional practice, with the highest scores reported in the Classroom Organization domain (Booker, 2014; Malmberg et al., 2010; Neergaard & Smith, 2012). However, two of these studies identified the greatest improvement in Classroom Organization scores (Booker, 2014; Malmberg et al., 2010), while another study identified the largest improvement (based on growth coefficients) in Emotional Support scores (Neergaard & Smith, 2012). One study identified an inverted U-shaped change in Emotional Support, suggesting an increase and subsequent decline in this domain with early teaching experience (Malmberg et al., 2010). Although all three of these studies identified Instructional Practices had the lowest scores, and least amount of growth during a teacher's early career years (Booker, 2014; Malmberg et al., 2010; Neergaard & Smith, 2012). Other researchers reported unclear (Caspersen & Raaen, 2015) or complex findings (Cherng & Halpin, 2016), suggesting other

contextual factors significantly impact the relationship between CLASS-S scores and teacher characteristics

Of the key themes, this review provided a more in-depth summary exploring the relationship between classroom practices (i.e., CLASS-S) and student outcomes. In brief, student outcomes fell into three broad themes: student achievement, student engagement, and student's perspectives of classroom context. Based on findings previously summarized, the relationship between CLASS-S scores and student achievement remains unclear. Specifically, the various methods used to measure student achievement (e.g., value-added scores, GPAs, state mandated subject or achievement tests) have received criticism (e.g., AERA, 2015; Kuncel, Credé, & Thomas, 2005). Despite challenges, researchers have described some emerging evidence of positive associations between the CLASS-S and student achievement.

Studies of student engagement primarily included more than one measure of engagement (i.e., student-reported, teacher-reported, and CLASS-S observations). Researchers identified varying patterns of engagement over the course of the school year, and suggested relationships with other contextual factors such as autonomy, peer relatedness, and motivation (e.g., Hafen et al., 2012; Ruzek et al., 2016). Overall, researchers found student engagement as observed on the CLASS-S protocol provided a strong indicator of student engagement (e.g., Virtanen, Lerkkanen, Poikkeus, & Kuorelahti, 2015), and supported its utility as an observational measure in addition

to more traditional student- or teacher-reported indicators. As such, researchers provided emerging support for the CLASS-S as a measure of student engagement.

In the last theme addressing the relationship between classroom and student outcomes, researchers used the CLASS-S to examine student's perspectives of classroom context, specifically classroom structure, and classroom climate. Researchers found complicated relationships between student ratings and observations (Spearman & Watt, 2013). For example, girls who rated a dimension as low and attended a class where the dimension received low observational ratings had lower levels of motivation for science (Spearman & Watt, 2013). However, girls who rated a dimension as high and were observed in a classroom where the dimension was also rated as high on the CLASS-S, motivation for science decreased from the start of the school year (Spearman & Watt, 2013). In contrast, researchers exploring multiple factors of classroom climate (i.e., academic and emotional peer support, teacher-student relationship, classroom misbehavior) found no associations between student-ratings and classroom observational ratings on the CLASS-S (Yoder, 2012). Therefore, the association between classroom observations, and student perspectives is still uncertain.

Additionally, researchers have noted other yet to be researched factors that might be associated with the CLASS-S. In particular, adolescents have different developmental needs than do students in elementary or pre-school. Accordingly, linking underlying skills or abilities

associated with being a successful young adult to classroom interactions might help to identify how best to support classroom environments for older students. For example, self-determination theory (Deci & Ryan, 2000) was cited as one of the underlying constructs informing the CLASS-S. Therefore, indicators of competence, relatedness, and autonomy should be explored in association with the CLASS-S. Typically, measures of student success in middle and high schools tend to be very achievement driven; yet other skills (e.g., higher-order thinking, self-regulation, executive functioning, social skills) have been overlooked in research linking measures to student outcomes (Burchinal, 2018).

Recommendation 2: Further development of key themes. As described previously, research exploring the association between the CLASS-S and key themes has identified both emerging and divergent findings. Therefore, more research is required to clarify findings across the eight themes identified in this review, and with other underlying constructs related to adolescent development. More work examining patterns of school/classroom differences, and research to account for other contextual factors contributing to these differences is warranted. For example, researchers (e.g., Bell et al., 2012) have suggested student factors contribute a significant amount to instructional quality, and therefore directly impact CLASS-S scores. Consequently, identifying impactful student factors in middle school and high school settings, might help teachers and administrators change practices to support teacher-student interactions in

higher grades. Furthermore, a better understanding of differences between teachers of varying subject matters might help target professional development for teachers by subject.

Caution should be noted inherent to the value-added models chosen by authors, as not enough is known about these models and their ability to predict performance or differentiate between schools. Researchers have suggested there was more within school, than between school variability based on the CLASS-S (Rutledge et al., 2012). If in fact value-added models identified schools where teachers contributed to greater gains in student achievement, it would follow that classrooms in HVA schools would have higher CLASS-S scores. Therefore, as recommended by the AERA (2015), more investment in alternative methods and models for evaluating educators is needed to better understand these models prior to relying on them to drive decisions, or draw conclusions.

In addition, further research exploring patterns of CLASS-S scores during early career teaching might provide insight into areas needed for additional instruction in teacher education programs, or for additional early career support or mentorship. In particular, teachers often scored lowest on the Instructional Supports domain, and authors have suggested novice teachers have difficulties articulating their support needs (Caspersen & Raaen, 2015), therefore explicitly teaching instructional practices based on teacher-student interactions in education programs, and providing early career mentorship, might lead to overall gains in effective teaching. Mapping

teacher characteristics in this review identified much has yet to be researched. For example, examining the CLASS-S with a sample of highly effective teachers, in longitudinal studies, or identifying other contextual factors such as the contribution of school climate, or student outcomes to CLASS-S scores, remain areas for future study.

Importantly, there remains much to be clarified with respect to relationships between classroom and student outcomes. For example, limited evidence exists linking middle or high school teacher-student interactions to any indicator of student achievement (Allen et al., 2013). Therefore, examining the relationship between the CLASS-S and student achievement (using a variety of measurement methods) remains a key area for further research. Bell and colleagues (2012) suggested alternate measures of student achievement be considered, and proposed other constructs including (but not limited to) measures of graduation rate, student portfolios, course talking patterns, and measures of student beliefs. Furthermore, research is needed to explore the potential for the Student Engagement domain to serve as a stand-alone observational measure of student engagement, and to investigate its relationship to other outcomes. Last, linking student perspectives of classroom environment, teacher-student interactions, and what they need to be successful, to the CLASS-S would strengthen our understanding of middle and high school environments.

Despite a growing research base supporting the use of the CLASS in early years and elementary school settings, researchers have yet to produce substantial evidence specific to the CLASS-S in middle and high school settings. However, the present review suggests the CLASS-S might be a valuable tool for exploring differences in middle and high school settings, and therefore calls for more research. In particular, studies examining differences between groups identified CLASS-S scores were higher in high or honors track classrooms than regular track classrooms, and lower in middle and high school grades than elementary grades. When considering there has been no clear pattern of differences identified between schools (i.e., low and high value-added schools, or public versus charter schools), it remains there might be other unidentified student-specific factors contributing to CLASS-S scores. Based on this overall pattern, there may be greater classroom variability than school-wide variability on the CLASS-S, suggesting that individual teachers can have a great impact in their classroom despite other identified school-based factors. Therefore, as suggested by the AERA (2015), more research is needed before alternative measures of teaching effectiveness are used in high stakes evaluations, or to drive decision-making processes at the school-wide level.

Finally, exploring the relationship between the CLASS-S and other underlying constructs or processes, such as higher-order thinking, self-regulation, executive functioning, social skills, and self-determination theory, are worthy of further research (Burchinal, 2018). In sum, research

linking the CLASS-S to theories and measures associated with constructs of adolescent development, and examining characteristics beyond student academic achievement would deepen our understanding of the context of middle and high school classrooms, and in turn strengthen our ability to support adolescents to gain skills to become successful young adults.

Gap 3: Ability to enhance classroom practices. One area identified in this review with potential to impact teaching practices is the development of teaching interventions based on the CLASS-S model. Still, there appears to be a practice-to-research gap, as practice seems to have outpaced research. One intervention, MTP-S, has been associated with two published research articles included in this review (i.e., Allen et al., 2011, Allen et al., 2015). According to the researchers, changes in professional practice have in turn been measured as gains in student achievement (Allen et al., 2011, Allen et al., 2015). However, clear links to increased classroom practices have not been well defined in the literature. Furthermore, anecdotally, there are other coaching and professional development initiatives based on the CLASS-S occurring in both the United States, and abroad.

Recommendation 3: Further research exploring the ability to enhance classroom practices. Because there is a likelihood that interventions based on the CLASS-S result in changes to teacher-student interactions, more research is needed to explore this impact. As more is understood about the relationships between the CLASS-S and other key constructs, these

relationships should also be explored in the context of a CLASS-S based intervention. Unanswered questions include (but are not limited to): do interventions impact other key constructs, how long are changes sustained, are there optimal grade(s) during the school years to have a high-quality teacher (e.g., pre-kindergarten, Grade 4, Grade 9), what are the best practices for coaching interventions (e.g., target subjects, grades, mentorship, booster sessions, methods of providing feedback). Furthermore, more research exploring the effects of explicitly teaching the CLASS-S model in pre-service teacher education programs is warranted.

Limitations

The present article was a systematic scoping review and followed recommended guidelines (Arksey & O'Malley, 2005; Peters et al., 2015). As noted previously, the intention was to provide an overview of published peer-reviewed research and grey literature, and not to provide quality ratings, effect sizes, or conduct secondary analyses. The results are limited and may not be generalized beyond a current mapping of the literature. Some additional limitations include the overall quality of the reporting standard in some of the articles reviewed. As such, information in some articles was unclear preventing, for example, retrieval of information to complete every field of the data extraction form. Specifically, identifying research methods, and making associations between article authors and the original developers of the CLASS-S was sometimes challenging. In addition, many articles stemmed from data obtained for larger

projects, and therefore limitations may be assumed related to drawing conclusions from the same sample.

Conclusion

This scoping review provided an overview of the CLASS-S in published literature, mapped key themes addressed by authors using the CLASS-S, presented findings about the relationship between classrooms (i.e., CLASS-S) and student outcomes, and identified gaps and summarized recommendations for the future use of the CLASS-S. In brief, 55 articles were identified for inclusion since 2009, and search results reflected the majority of these studies originated from the United States. Most included studies used earlier iterations of the protocol with only approximately 4% using the final published version of the CLASS-S. Furthermore, many studies drew data from larger projects, limiting the number of unique data sets used across articles. Despite some agreement about reporting training and coding procedures, the process of double coding and reporting inter-rater reliability statistics varied widely across articles.

Key constructs mapped across articles fell into eight broad categories (see Figure 2.4) including the CLASS-S model, CLASS-S procedures, relationships between classroom and student outcomes, school/classroom contrasts, teacher characteristics, teacher/curricular interventions, teacher education program assessments, and case studies. Of these key themes, the relationships between classroom (i.e., CLASS-S) and student outcomes were presented more

completely in this review. Specifically, three themes summarizing findings specific to linking the CLASS-S to student achievement, student engagement, and student perspectives were described. In sum, authors described some emerging and conflicting findings linking the CLASS-S to student achievement. Authors provided some evidence the CLASS-S student engagement measure contributes to our understanding of student engagement beyond other indicators (e.g., teacher or student reports). Furthermore, authors explored the relationship between student perspectives and the CLASS-S and reported limited findings.

As a scoping review (Grant & Booth 2009), the purpose of the present study was to review the current body of emerging literature using the CLASS-S. Scoping reviews assist policymakers to decide if further systematic reviews are needed, and are particularly helpful to summarize findings when more precise statistical methods, or quality assessments, are not yet applicable (Grant & Booth; Peters et al., 2015). In summary, the present scoping review examined the use of the CLASS-S in the published literature and identified gaps in knowledge about its psychometric properties, its association with key themes (e.g., student achievement), and its ability to enhance classroom practices. Briefly, research to clarify double coding and inter-rater reliability procedures, to provide further evidence to support the validity of the CLASS-S for each of its potential uses, to broaden its association with key themes, and to expand on its potential to guide interventions and pre-service teacher education programs is

needed. Once additional research on these topics has been published, a systematic review may be warranted. As it currently stands, the CLASS-S is emerging as a useful and valuable tool, but more evidence-based support is needed. Research using the CLASS-S may help deepen our understanding of classroom environments and supports needed to strengthen these environments for middle and high school students, a population commonly overlooked in educational research.

Table 2.1.

Versions of the CLASS-S (domains and dimensions) as identified in published research.

	Version 1 (n = 26)	**Version 2** (n = 4)	**Version 3** (n = 16)	**Version 4** (n = 8)	**Version 5** (n = 2)
Emotional Support					
Positive Climate	**x**	**x**	**x**	**x**	**x**
Teacher Sensitivity	**x**	**x**	**x**	**x**	**x**
Regard for Adolescent Perspectives	**x**	**x**	**x**	**x**	**x**
Negative Climate	**x**	**x**	**x**		
Classroom Organization					
Behavior Management	**x**	**x**	**x**	**x**	**x**
Productivity	**x**	**x**	**x**	**x**	**x**
Instructional Learning Formats	**x**	**x**	**x**		
Negative Climate				**x**	**x**
Instructional Support					
Content Understanding	**x**	**x**	**x**	**x**	**x**
Quality of Feedback	**x**	**x**	**x**	**x**	**x**
Analysis and *Problem Solving*	**x**	**x**	**x**	**x**	
Analysis and *Inquiry*					**x**
Instructional Learning Formats				**x**	**x**
Instructional Dialogue			**x**		**x**
Procedures & Skills		**x**			
Student Engagement	**x**	**x**	**x**	**x**	**x**

Note: Dimensions that changed across versions are italicized.

Table 2.2.

Characteristics of included publications

	Author(s), year	Research Design	Larger Project	Teachers *N*	Student *N*	Grades	Key Themes	Findings Specific to CLASS-S
1	Allen, Hafen, Gregory, Mikami, & Pianta, 2015	Experimental	MTP-S (2)	86	1194	11% middle school, 89% high school	Relationship between classroom and student outcomes, Teacher/curricular interventions	• Students made significant gains on a measure of academic achievement when their teacher participated in MTP-S, when compared to students of teachers who did not participate. • Gains were reported by the researchers to translate to an increase on achievement tests from the 50th percentile to the 59th percentile for the average student.
2	Allen, Pianta, Gregory, Mikami, & Lun, 2011	Experimental	MTP-S	78	2237	Mean grade = 8.06, SD = 1.5	Relationship between classroom and student outcomes, Teacher/curricular interventions	• Students made significant gains on a measure of academic achievement when their teacher participated in MTP-S, when compared to students of teachers who did not participate. • Gains were reported by the researchers to translate to an increase on achievement tests from the 50th percentile to the 59th percentile for the average student.
3	Allen et al., 2013	Experimental	MTP-S	37	643	Mean grade = 8.3, SD = 1.5	Relationship between classroom and student outcomes, CLASS-S model	• Scores on the Positive Climate, Regard for Adolescent Perspectives, Instructional Learning Formats, and Analysis and Problem-Solving dimensions were predictive of greater student achievement. • Confirmatory factor analysis supported the three-factor model proposed by the authors.
4	Amato, 2012	Case Study	-	4	4	High school	Case study	• The Regard for Adolescent Perspective dimension of the CLASS-S was used to supplement qualitative observations of the classroom experience of four English

								Language Learning students in a high school environment. The only explicit findings based on the CLASS-S provided were scores ranged from 2-6 on Regard for Adolescent Perspective.
5	Athanases et al., 2016	Case Study	UCA	6	-	High school (grades 9-12)	Case study	• CLASS-S was used as an observational measure, and written notes used for scoring were included in qualitative analyses. Researchers identified teachers provided indicators of Positive Climate, Teacher Sensitivity, and Behaviour Management. However, researchers identified a need for increased academic rigor, (i.e., Instructional Practices on the CLASS-S), to deepen instruction in order to reflect a college-bound culture. For example, the researchers distinguished between talking about college, to engaging in college-level discourse.
6	Athanases & de Oliveira, 2014	Case Study	UCA	2	48	54% grade 11, 46% grade 12	Case study	• CLASS-S was reported by the authors to provide additional information about classroom practices in a case study examining two teachers' approaches to scaffolding for Latina/o English language learners. Limited scores were reported on a few dimensions for one teacher, and for only one dimension (Quality of Feedback) for the second teacher.
7	Bell et al., 2012	Exploratory Research	TUCC	82	-	32% middle school, 68% high school	Relationship between classroom and student outcomes, CLASS-S model, CLASS-S procedure	• Researchers reported evidence linking the CLASS-S to student achievement was divergent. • Unexplained errors were identified by the researchers as a large source of variation among CLASS-S scores. • The researchers suggested a large number of observations across time are required

								prior to making any conclusions about teaching.
8	Bell et al., 2014	Mixed Methods	MET and UTQ	MET: nearly 3000; UTQ: 458	-	Grades 6, 7, and 8	CLASS-S procedure	• Observers displayed the strongest pattern of agreement and accuracy in the Classroom Organization domain (note: overall agreement metrics were stronger in the MET data as compared to the UTQ data). Qualitatively, observers noted that Classroom Organization was easier to score, and Instructional Support was more difficult as it required higher-inference judgments.
9	Booker, 2014	Longitudinal	AIM	62	-	-	Teacher characteristics	• Beginning math teachers were predicted to improve significantly over a three-year time period across all domains of the CLASS-S, with the largest growth reported in the Classroom Organization domain.
10	Brandt, Meyers, & Molefe, 2013	Experimental	eMINTS	191	3610	All grades, emphasis on grades 7 and 8	Teacher/curricular interventions	• After one year of participation in the eMINTS program, the researchers noted that Grade 7 and 8 teachers were observed to use more instructional practices to support inquiry-based learning. Explicit CLASS-S findings were not reported.
11	Casabianca, Lockwood, & McCaffrey, 2015	Longitudinal	UTQ	458	-	34% grade 6, 29% grade 7, 36% grade 8, rest from mixed grades	CLASS-S procedure	• Researchers explored trends in classroom observation scores and identified rater effects (i.e., rater error and residual error) as the largest source of variance. Rater drift was estimated to increase during the study. • To reduce errors, the authors suggested providing raters with field experience as part of their training, and/or to remove struggling raters when identified through calibration exercises.

12	Casabianca et al., 2013	Exploratory Research	TUCC	82	-	32% middle school, 68% high school	CLASS-S procedure	• Researchers identified some minor variation between live and videotaped scoring in their sample, and noted that both methods included large errors and low reliability. • Including a large number of ratings on multiple classroom sessions improved reliability and reduced error beyond the inclusion of live scoring.
13	Caspersen & Raaen, 2014	Case Study	-	8	-	50% lower secondary schools (grades 7-10), and 50% elementary schools (grades 1-6)	Teacher characteristics, CLASS-S model	• Researchers observed and rated 8 novice teachers using the CLASS-S, then reviewed the video footage with the teacher while conducting interviews based on the CLASS-S model. Specifically, researchers were seeking the opinions of novice teachers on the relevance of the CLASS-S constructs to their teaching practice, and if they had learned them in pre-service teaching programs. CLASS-S findings were unclear, overall findings reported by the researchers were that novice teachers may have difficulty articulating their needs and this may be a barrier to receiving support from superiors and colleagues.
14	Cherng & Halpin, 2016	Secondary Data Analysis	MET	1,680	51,347 (1st year MET)	Grades 6-9	Teacher characteristics	• Researchers used the results of the CLASS-S to create an aggregated indicator of teacher effectiveness. In hierarchical linear regression modeling, the researchers suggested minority students favored minority teachers, and reported students' perceptions were influenced by their academic performance, teacher characteristics, the teaching conditions (i.e., as reported by teachers), and overall teacher effectiveness.

15	Culp, Martin, Clements, & Presser, 2015	Experimental	-	41	914	Mean age = 12.7 (SD = 0.5)	Relationship between classroom and student outcomes, Teacher/curricular interventions	• The researchers reported an interaction effect between CLASS-S scores and student outcomes for the classrooms in the treatment group. Researchers stated teachers who scored higher on the CLASS-S and used the Exploring Photosynthesis program had better student outcomes. The authors suggested more effective teachers were able to implement technology to enhance learning, by making meaningful connections for students.
16	Donaldson, LeChasseur, & Mayer, 2017	Mixed Methods	-	149	-	High school	School/classroom contrasts	• Using multilevel modeling, scores on the CLASS-S were significantly different between low and high track classrooms, with low track classrooms experiencing less Emotional, Organizational, and Instructional Support. • Using multilevel modeling, Content Understanding and Analysis and Problem Solving scores were higher in Mathematics classes when compared to English classes.
17	Durksen et al., 2017	Case Study	-	6	-	Upper primary and secondary schools	Case Study	• The researchers noted that the CLASS-S helped them focus on observable teacher-student interactions in the classroom, and they integrated their results by categorizing their coded transcripts using CLASS-S domains. Overall, teachers reported motivational or engagement strategies consistent with the domain of Classroom Organization. The researchers noted increasing positive motivation or reducing negative motivation was facilitated through the use of interactions that were high in Emotional and Instructional Support. They also identified increasing positive engagement or reducing negative engagement often

								involved teacher practices consistent with Instructional Support and effective Classroom Organization.
18	Friedman, 2014	Exploratory Research	MTP-S (2)	74	-	Middle and High schools	Teacher characteristics	• Teachers with higher scores on a composite of the Teacher Emotional Intelligence Measure (TEIM) also had higher scores on the dimension, Regard for Adolescent Perspectives on the CLASS-S. • The author posited that the composite TEIM indicated higher emotional intelligence, and therefore a greater ability to recognize and meet the developmental needs of adolescents in the classroom.
19	Gamlem & Munthe, 2014	Descriptive Research	-	28	-	Grades 8-10	CLASS-S model	• Using the CLASS-S and a modified version of the Quality of Feedback dimension, researchers found the Quality of Feedback dimension emphasized encouragement-, over learning-oriented feedback.
20	Gill et al., 2016	Secondary Data Analysis	MET	662	-	Grades 4-9	CLASS-S model	• Researchers compared observation protocols used in the MET (CLASS, FFT, MQI, PLATO, and Uteach) and identified and renamed10 common themes across all protocols. The CLASS captured 9 of the 10 overarching themes (classroom management, supportive learning environments, student focus, active student participation in class activities, student intellectual engagement with content, lesson structure and facilitation, content understanding, language and discourse, feedback and assessment). The only theme not represented on the CLASS protocol, was an indicator of teacher professionalism. • Specific to the MET project data, researchers identified teacher observation

								scores were significantly impacted by the percentage of racial/ethnic minority students enrolled, when observations occurred in English Language Arts classrooms. However, the impact of student outcomes was lessened (<5% of the 64 regressions performed) when observations were conducted using the CLASS when compared to observations using the Frameworks for Teaching protocol (25% of the 48 regressions performed).
21	Gitomer et al., 2014	Exploratory Research	TUCC	82	-	50% middle school, 50% high school	CLASS-S procedure	• Researchers indicated all scores fell within narrow ranges and observers became more consistent over time (and agreed most on Classroom Organization domain). Instructional Support was the most difficult domain to score accurately and consistently.
22	Gregory et al., 2015	Experimental	MTP-S (2)	82	979	Middle and high schools; Mean = grade 10	Teacher characteristics, Teacher/curricular interventions	• Teachers referred fewer students for exclusionary discipline (i.e., to the principal's office) after participating in MTP-S, particularly fewer African American students.
23	Gregory et al., 2014	Experimental	MTP-S	87	1669	61% middle schools, 39% high schools	Teacher/curricular interventions	• A modest increase in student engagement was identified in classrooms where the teacher participated in MTP-S, mediated by scores on the Analysis and Problem Solving (or Inquiry) and Instructional Learning Formats dimensions.
24	Gregory et al., 2016	Experimental	MTP-S (2)	Year 2 = 86; Year 3 = 79	Year 1: 1195; Year 2: 1163	11% middle school, 89% high school	Teacher/curricular interventions	• The reduction in referrals of African American students was maintained across two years of coaching, and in the year following the conclusion of the intervention. • Further analysis highlighted improvements in classroom practices,

								particularly those related to problem solving and higher level thinking (i.e., the Analysis and Inquiry dimension), as key factors contributing to the effectiveness of MTP-S.
25	Grossman, Loeb, Cohen, & Wyckoff, 2013	Exploratory Research	-	24	-	Grades 6-8	School/classroom contrasts	• Productivity, Student Engagement, and Behavior Management were highly correlated (at least .94), the correlations made it difficult to assess whether particular instructional elements were individually related to higher value-added scores or if they were tapping into a broader underlying dimension that characterized a difference between the groups.
26	Hafen et al., 2012	Exploratory Research	MTP-S	34	578	Grades 9-12 (Mean = 10.24, SD = 0.88)	Relationship between classroom and student outcomes,	• Levels of student-reported and observed CLASS-S Student Engagement were predicted by classroom-level perceptions of autonomy at the beginning of the year. The measures of student-reported and observed engagement were not correlated at the beginning of the year, however were strongly associated by the end of the year. The authors suggested observed engagement (at the beginning of the year in particular), was a stronger measure than student self-reports.
27	Hafen, Hamre, et al., 2015	Secondary Data Analysis	MTP-S, MET, TUCC, & UTQ	MTP-S: 67; MET: 875; TUCC: 82; UTQ: 458	-	MTP-S: Grades 6-11; MET: Grades 6-8; TUCC: 50% middle schools, 50% high schools;	CLASS-S model	• A revised factor model (where Negative Climate was moved to the Classroom Organization domain and Instructional Learning Formats was associated with the Instructional Support domain) produced a better fit. Findings led to the final structure of the published CLASS-S measure.

						UTQ: middle schools		
28	Hafen, Ruzek, Gregory, Allen, & Mikami, 2015	Experimental	MTP-S (2)	86	1195	11% middle school, 89% high school	Teacher/curricular interventions	• Participating teachers predicted better student future educational attainment despite the level of behavior problems their students self-reported at the beginning of the year.
29	Halpin, & Kieffer, 2015	Secondary Data Analysis	MET	381	-	Grades 6-8	CLASS-S model	• Researchers proposed a method of latent class analysis and identified 4 distinct profiles of instructional practice that demonstrate the statistical relationship among items, as an illustration of a potential method for future research. The authors reported their method better represented the complexity of the relationships among dimension scores, in contrast to typical rank or item level findings.
30	Jamil, Downer, & Pianta, 2012	Exploratory Research	-	509	-	-	Teacher education program assessments	• Pre-service teachers' ratings on the CLASS-S were not related to their beliefs of teaching self-efficacy at the end of the teaching preparation program.
31	Joe, McClellan, & Holtzman, 2014	Secondary Data Analysis	MET	50	-	-	CLASS-S procedure	• Researchers found strong correlations between segment and whole scores, particularly scores from middle segments of a classroom session, therefore middle segment scores approximated the score of the full session.

32	Kane & Staiger, 2012	Exploratory Research	MET	1333	44500 (total MET project)	Grades 4-8	Relationship between classroom and student outcomes, CLASS-S model, CLASS-S procedure	• The CLASS was modestly related to gains in student achievement when used alone, and more strongly correlated when combined with student feedback. • Findings from the CLASS were highly correlated (0.88) with another classroom observation protocol, the Framework for Teaching. • Sources of variance in scores were identified across CLASS domains as follows: 23-32% attributed to differences among teachers, 18-28% attributed to between-lesson variance for teachers, 0-3% attributed to course-section variation (when teachers taught more than one section of the same course), 10-11% attributed to rater effects (patterns of high or low rating), and 32-42% attributed to residual variance. • The highest scores on the CLASS were received on Behavior Management and Productivity dimensions (85% scored 5 or higher), and less than 1% scored 5 or higher on the Negative Climate dimension. While fewer than 10% scored 6 or 7 on the dimensions of Regard for Student Perspectives, Quality of Feedback, Instructional Dialogue, and Content Understanding. The Analysis and Problem-Solving dimension saw only 20% scoring between 4-7, meaning that 80% of scores fell in the 1-3 range. • High quality observations (i.e., more reliable scores) require clear standards, trained and certified observers, and multiple observations per classroom.

33	Malmberg & Hagger, 2009	Exploratory Research	-	30	-	High school	Teacher education program assessments	• Higher Emotional Support and Student Engagement scores predicted higher instructional agency at the end of a teacher education program.
34	Malmberg, Hagger, Burn, Mutton, & Colls, 2010	Longitudinal	-	17	-	Grades 7-12	CLASS-S model, Relationship between classroom and student outcomes, Teacher characteristics	• Confirmatory factor analysis supported an a priori reduction of dimensions per domain. Emotional Support included Positive Climate and Regard for Adolescents, Classroom Organization included Behavior Management and Productivity, and Instructional Support included Content Understanding, Analysis and Problem Solving, and Quality of Feedback. • Researchers observed teachers during their pre-service education program, and in the first two years of their teaching careers. They identified that Classroom Organization increased linearly over time, while Emotional Support displayed an inverted U-shape over time. For older students, Emotional Support was lower. • Larger class size was related to increased Student Engagement. Instructional Support was not related to Student Engagement.
35	Mashburn, Meyer, Allen, & Pianta, 2014	Experimental, Secondary Analysis	MTP-S	47	1366	Grades 6-12	CLASS-S procedure	• The researchers identified that 2 20-minute video-recorded segments presented in a random order for coding was the most statistically sound.
36	McCaffrey, Yuan, Savitsky, Lockwood, & Edelen, 2015	Secondary Data Analysis	UTQ	458	-	Grades 6-8	CLASS-S model	• The authors identified scoring errors between raters on the CLASS-S contributed to differing factor structures at the teacher level. Furthermore, averaging scores contributed to a different factor structure. Applying alternative hierarchical estimation approaches, the authors identified a two-factor model (strongly

								correlated). Authors therefore proposed a one-factor model might be adequate. The researchers suggested an overall CLASS-S score based on multiple observations might be the best indicator of effective teaching.
37	Meyer, 2011	Descriptive Research	-	41	-	High school	-	• The author reported scores from observations conducted once per classroom and described the means and frequencies of each dimension, no further CLASS-S findings were provided.
38	Meyers, Molefe, Brandt, Zhu, & Dhillon, 2016	Experimental	eMINTS	200	3072	Grade 7 and 8	Teacher/curricular interventions	• After two years of participation in the eMINTS program, the researchers used CLASS-S scores to measure two of their strategies "community of learners" and "inquiry-based learning" by placing outcomes on a Rasch scale for comparison. • The authors reported positive significant results with a medium effect size, indicating that the eMINTS program led to gains in community of learners, and inquiry-based learning for both their treatment groups when compared to the control group.
39	Mihaly & McCaffrey, 2014	Secondary Data Analysis	MET	-	-	Grades 4-8	Relationship between classroom and student outcomes, School/classroom contrasts	• Teachers in Grades 4 and 5 scored higher than Grade 6 teachers, and Grade 6 teachers scored higher than teachers in Grades 7 and 8. • Grade 7 and 8 mathematics teachers scored lower than English language teachers. • Classroom composition (i.e., prior year achievement scores, student average age, student socio-economic status) accounted for a small portion (i.e., approximately 30 percent) of the magnitude of the difference between scores.

								• No grade-level differences were found in the ability of the CLASS to predict student achievement gains.
40	Neergaard & Smith, 2012	Longitudinal	AIM	62	-	-	Teacher characteristics	• With regards to the CLASS-S observed in the final year of a teachers education program and first two years of professional practice, the authors reported highest scores on Classroom Organization and lowest on Instructional Support, across all time points. Based on the growth coefficients, scores on each of the CLASS-S domains significantly improved over time. Emotional Support improved by the largest degree (a fourth of a point), while Classroom Organization improved by a smaller margin (a third of point), and Instructional Support improved the least.
41	Ozer & Douglas, 2015	Exploratory Research, Mixed Methods	-	4	-	Aged 14-19 years (Mean = 16.8, SD = 1)	-	• Despite describing the use of the CLASS-S as a part of the assessment of classroom settings in the methods section of the article, no further findings from the measure are included.
42	Park, Chen, & Holtzman, 2014	Secondary Data Analysis	MET	-	-	-	CLASS-S procedure	• The authors found no impact of coder background on scoring accuracy. Coders who were familiar with the CLASS measure prior to the study, who noted they reflected on their own teaching practice, and noted the tool could be helpful in professional development, were more accurate in their scoring. The authors also found no impact of classroom setting/composition on scoring accuracy. • Researchers made recommendations for future research utilizing classroom observations: 1) ensure the protocol trains and monitors coder performance, 2) ensure ongoing statistical monitoring of coders,

								and 3) give individual feedback and additional training if needed.
43	Ripski, LoCasale-Crouch, & Decker, 2011	Longitudinal	-	41	-	K-12	Teacher education program assessments	• Pre-service teachers who reported higher ratings of extraversion and depression at the beginning of a teacher education program, scored lower on Instructional Support in the last year of their education program.
44	Rutledge, Cohen-Vogel, & Osborne-Lampkin, 2012	Longitudinal Case Study	NCSU	73	-	Grades 9-12 (primarily 10th grade)	School/classroom contrasts	• Authors noted no major differences on the CLASS-S when comparing low to high value-added schools. • When comparing CLASS-S scores across four case study schools, researchers used multilevel modeling to identify Emotional Support was lower in regular track classes when compared to honors classes (particularly in one school). According to the researchers, results highlighted the presence of more within-school variability as compared to between-school variability.
45	Ruzek et al., 2016	Exploratory Research	MTP-S	68	960	63% middle school, 38% high school	Relationship between classroom and student outcomes	• The researchers reported classrooms with higher Emotional Support scores on the CLASS-S in the fall, had students who reported increased behavioral engagement and mastery motivation at the end of the school year. Students also reported more autonomy and positive relationships with their peers at the mid-point of the year when they were in classrooms with higher fall observed Emotional Support. The researchers noted that the indirect effect of Emotional Support on engagement and motivation was mediated by autonomy and peer relatedness.

46	Ruzek, Hafen, Hamre, & Pianta, 2014	Secondary Data Analysis	MET	1580	-	Grades 4-9	Relationship between classroom and student outcomes, CLASS-S model, School/classroom contrasts	• Based on spline regression methods used to identify active ranges and CLASS score thresholds, researchers suggested the relationship between classroom quality and student achievement was nonlinear, particularly when the overall CLASS score falls in the mid-range (e.g., 2.91-4.5). • Authors identified four strategies used to group teachers by effectiveness using the CLASS and Value-Added Modeling (VAM); 1) linear relationship, 2) CLASS manual categories, 3) CLASS distribution categories, 4) establish CLASS score thresholds. Strategy one identified a significant association, where a 1-point increase in CLASS score resulted in a 0.18 point gain in VAM. Strategy two identified the majority of teachers fell in the mid range (98.7%) with the remaining teachers falling in the low range (and no teachers in the high range). Strategy three grouped teachers in the bottom 10% (ineffective), next 40% (developing effectiveness), next 40% (effective), and top 10% (highly effective). Strategy four utilized simple regression to identify active ranges and thresholds, findings suggest that part of the relationship between classroom quality and student achievement is nonlinear, particularly when overall CLASS scores fall between 2.91 and 4.5.
47	Savitsky & McCaffrey, 2014	Secondary Data Analysis	TUCC	82	-	32% middle school, 68% high school	CLASS-S model	• Using a modeling approach described by the authors, data from the TUCC study were analyzed and a one-factor structure was identified as the best fit to examine teacher-level effects while controlling for all other levels in hierarchical data.

48	Shell, Gazelle, & Faldowski, 2014	Longitudinal	-	-	688	Elementary and middle school	School/classroom contrasts	• The researchers compared average scores on the Emotional Support domain from elementary (using the CLASS) to middle school (using the CLASS-S) and found the mean in middle school was significantly lower.
49	Smith, Cannata, & Haynes, 2016	Case Study	NCSU	72	-	Grades 9 and 10	School/classroom contrasts	• Researchers reported lower overall implementation of strategies to develop higher order thinking skills in the classroom, as captured by the Content Understanding and Analysis and Problem Solving dimensions of the CLASS-S. • HVA schools scored significantly higher on the dimensions of Emotional Support, except for Positive Climate, than LVA schools. • Researchers identified higher student engagement in two HVA schools and lower student engagement in one LVA school, while the second LVA school's student engagement score fell near the mean.
50	Smith, Preston, Haynes, & Neergaard Booker, 2015	Case Study	NCSU	73	-	Grades 9-12 (primarily 10th grade)	School/classroom contrasts	• The researchers did not identify any statistical differences between low and high value-added schools, however CLASS-S scores were higher in honors as compared to regular track courses.
51	Spearman, & Watt, 2013	Exploratory Research, Longitudinal	-	4	52	42% grade 7, 58% grade 8; aged 11-14 years (Mean = 12.79, SD = 0.75)	Relationship between classroom and student outcomes	• Researchers reported Positive Climate, Negative Climate, Regard for Adolescent Perspectives, Behavior Management, Quality of Feedback, and Student Engagement all impacted a student's perception of structure in the classroom. The researchers found girls, who rated a dimension as low and attended a class where the dimension received low ratings by CLASS-S observers, had lower levels

								of motivation for science. However, girls who rated a dimension as high and were observed in a classroom where the dimension was also rated as high on the CLASS-S, motivation decreased from the start of the school year. The researchers suggested this finding might be related to a ceiling effect. The researchers also found that girls who rated dimensions highly at the start of the year, yet were in classrooms where it was observed to be lower on the CLASS-S, reported the highest levels of motivation at the end of the year.
52	Virtanen, Lerkkanen, Poikkeus, & Kuorelahti, 2015	Exploratory Research	-	9 (8 teachers)	181	23% grade 7, 35% grade 8, 42% grade 9	CLASS-S model, Relationship between classroom and student outcomes	• Researchers identified variation in classroom quality on the CLASS-S in Finland, despite limited school-level variation in student outcomes. • The researchers identified higher quality classrooms, as observed on the CLASS-S, were associated with higher student engagement (as captured by student-ratings, teacher-ratings, and observations). Of the domains, Classroom Organization was associated with all three measures of engagement, while Instructional Support was associated with student-ratings and observations (but not teacher ratings). Emotional Support had an indirect effect on student engagement, via Classroom Organization and Instructional Support.
53	Wilson et al., 2015	Pilot Study	-	11	102	Aged 13-19 years (Mean = 15.83, SD = 1.26)	Teacher/curricular interventions	• CLASS-S scores significantly increased from pre- to post-intervention for the Productivity dimension. Scores also increased on the Quality of Feedback dimension, however the increase was not

								significant following procedure to ensure statistical rigor (i.e., Bonferroni correction).
54	Yoder, 2013	Mixed Methods	-	19	228	20% grade 4, 27% grade 5, 17% grade 6, 17% grade 7, 18% grade 8; Aged 9-14 years (Mean = 11.03, SD = 1.43)	Relationship between classroom and student outcomes, School/classroom contrasts	• Observer ratings on the CLASS-S were not related to any student outcomes (i.e., classroom climate, behavioral engagement, academic achievement) included in the study and the researcher suggested the results were due to the limited sample size. • The researcher did not find any significant differences on CLASS-S observation scores between one public school and one charter school. However, observer-reported Emotional Support scores predicted higher student-reported scores when students were asked to rate each of the CLASS-S dimensions.
55	Yuan, McCaffrey, & Savitsky, 2013	Secondary Data Analysis	UTQ and TUCC	TUCC: 82; UTQ: 458	-	-	CLASS-S model	• The authors proposed the CLASS factor structure varied by level, at the teacher level a one-factor model fit best, at the section level (for UTQ data) and session levels (for UTQ and TUCC data) there were no clear factor structures, and a two-factor model of Classroom Organization and "Teaching Capacity" (which included all other factors) fit at the rating level. The rating level was the score provided by one rater across one session. The authors suggested the rating process itself introduced structure into the scores, but that the data did not support a three-factor model.

Note: AIM = Assessing Induction and Mentoring, MET = Measures of Effective Teaching, MTP-S = My Teaching Partner – Secondary (MTP-S), MTP-S (2) = My Teaching Partner – Secondary replication study, NCSU = The National Center on Scaling up Effective Schools, TUCC = Toward an Understanding of Classroom Context, UCA = Urban College Academy, UTQ = Understanding Teaching Quality; HVA = High Value-Added, LVA = Low Value-Added

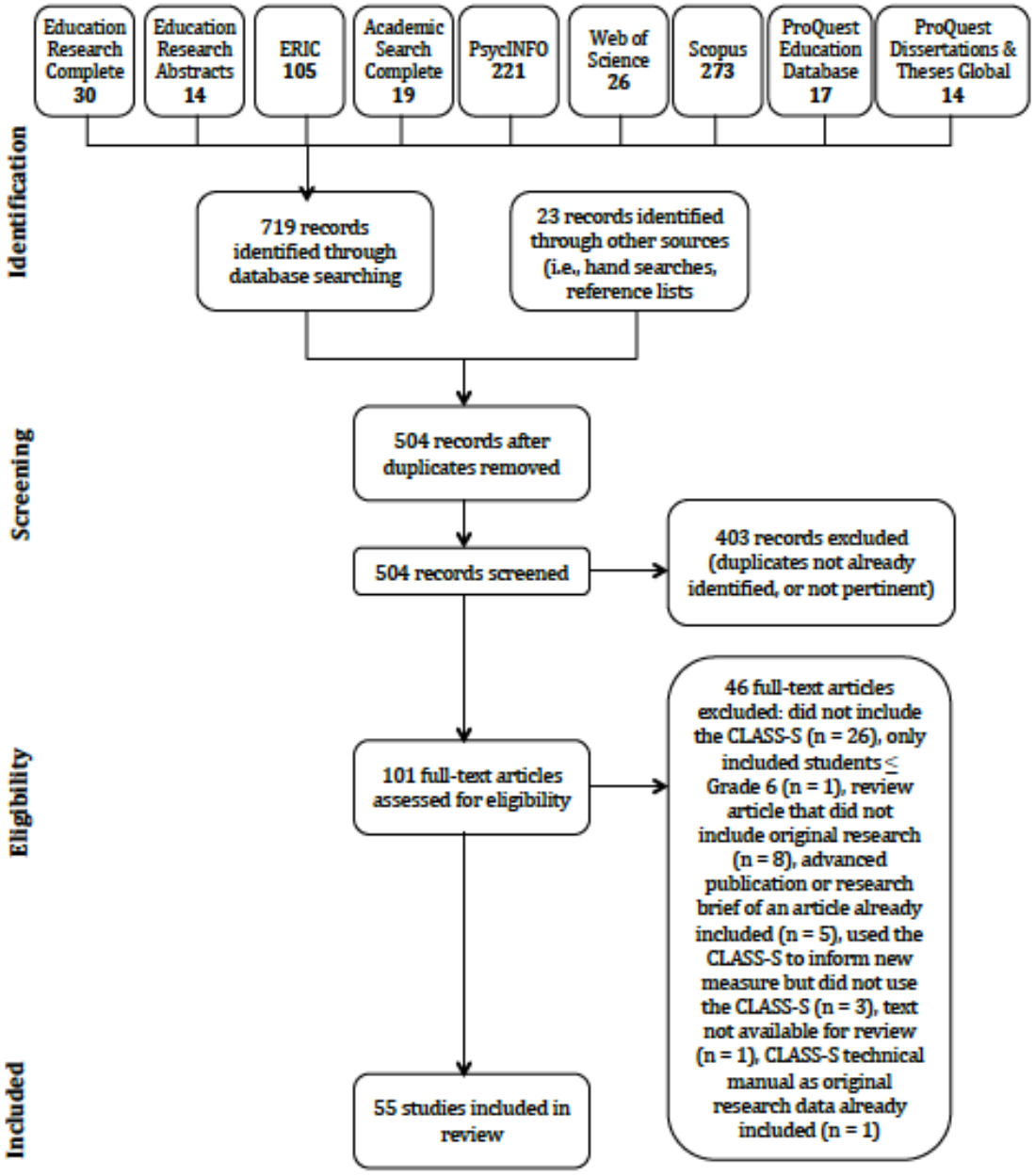


Figure 2.1. Flow diagram of the search strategy used to identify articles for review and inclusion.

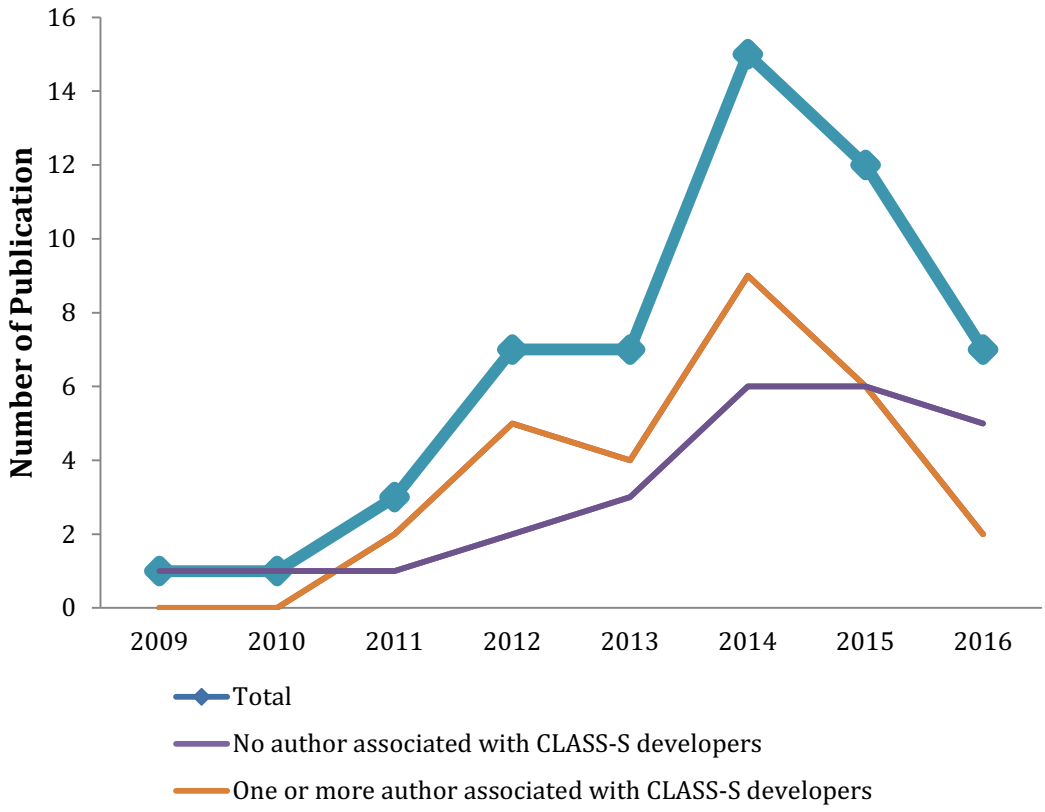


Figure 2.2. Number of CLASS-S publications identified for inclusion, total presented by year, and by association with protocol developers.

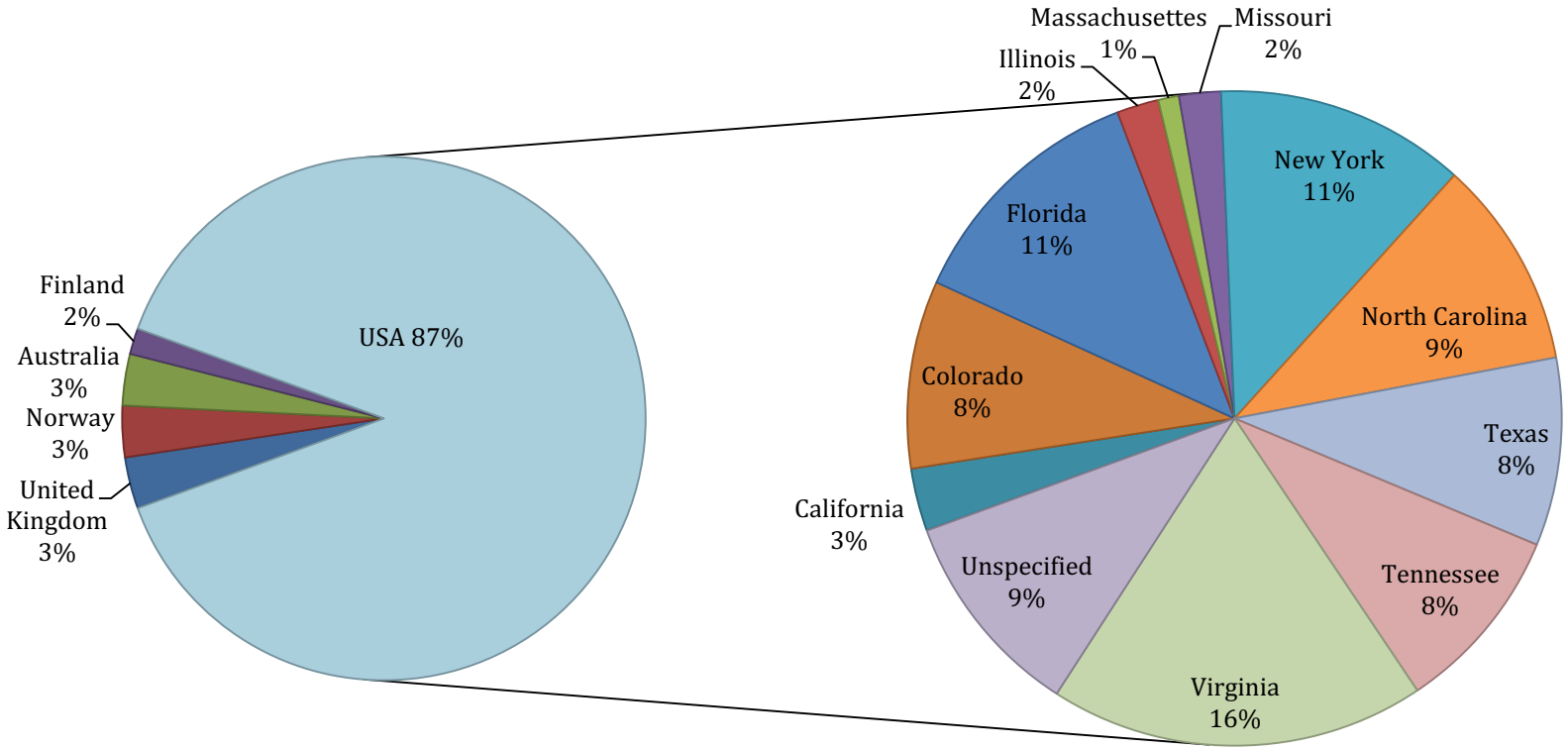


Figure 2.3. Location of study (or original data collection)

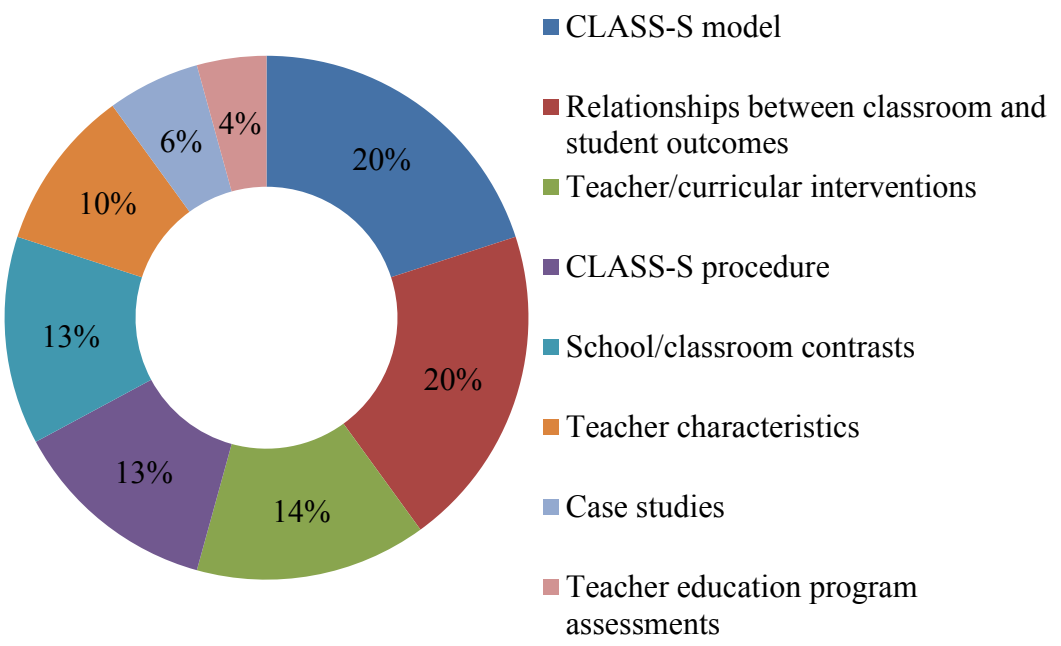


Figure 2.4. Chart by key theme, or reported finding(s) of published research included in the review. *Please note: Some articles included data that fell in more than one category.*

References

References marked with an asterisk (*) indicate studies included in the systematic scoping review.

*Allen, J. P., Hafen, C. A., Gregory, A. C., Mikami, A. Y., & Pianta, R. (2015). Enhancing secondary school instruction and student achievement: Replication and extension of the My Teaching Partner-Secondary intervention. *Journal of Research on Educational Effectiveness, 8*(4), 475-489. doi: 10.1080/19345747.2015.1017680

*Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333*(6045), 1034-1037. doi:10.1126/science.1207998

*Allen, J., Gregory, A., Mikami, A., Lun, J. Hamre, B. & Pianta, R. (2013). Observations of effective teacher-student interactions in secondary school classrooms: Predicting student achievement with the Classroom Assessment Scoring System-Secondary. *School Psychology Review, 42*(1), 76-97.

*Amato, W. W. (2012). *English language learners' experiences in mixed-language, mixed-culture classrooms* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Global. (UMI Number: 3525023)

American Educational Research Association. (2015). AERA statement on use of value-added

models (VAM) for the evaluation of educators and educator preparation programs. Educational Researcher, 44(8), 448-452. doi: 10.3102/0013189X15618385

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). AERA, APA, & NCME. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Archer, J., Kerr, K. A., & Pianta, R. C. (2014). Why measure effective teaching? In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), Designing teacher evaluation systems: New guidelines from the Measures of Effective Teaching project (pp. 415-443). San Francisco: Josey-Bass. Retrieved from http://k12education.gatesfoundation.org/wp-content/uploads/2015/11/Designing-Teacher-Evaluation-Systems_freePDF.pdf

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology, 8*(1), 19-32. doi: 10.1080/1364557032000119616

*Athanases, S. Z., Achinstein, B., Curry, M. W., & Ogawa, R. T. (2016). The promise and limitations of a college going culture: Toward cultures of engaged learning for low-SES Latina/o Youth. *Teachers College Record, 118*, 070306.

*Athanases, S. Z., & de Oliveira, L. C. (2014). Scaffolding versus routine support for Latina/o

youth in an urban school: Tensions in building toward disciplinary literacy. *Journal of Literacy Research, 46*(2), 263-299. doi: 10.1177/1086296X14535328

Battistich, V., Schaps, E., & Wilson, N. (2004). Effects of an elementary school intervention on students' "connectedness" to school and social adjustment during middle school. *Journal of Primary Prevention, 24*, 243–262. doi:10.1023/B:JOPP.0000018048.38517.cd

*Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*, 62-87. doi:10.1080/10627197.2012.715014

*Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., McCaffrey, D. F., Gitomer, D. H., & Pianta, R. C. (2014). Improving observational score quality: Challenges in observer thinking. In T. J. Kane, K. A. Kerr & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project* (pp. 50-97). Retrieved from http://janbri.nl/wp-content/uploads/2015/01/Designing-Teacher-Evaluation-Systems_freePDF.pdf

*Booker, L. N. (2014). *Examining the development of beginning middle school math teachers' practices and their relationship with the teachers' effectiveness* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Global. Vanderbilt University, Tennesee.

*Brandt, C., Meyers, C., & Molefe, A. (2013). *The impact of eMINTS professional development on teacher instruction and student achievement: Year 1 Report.* Retrieved from eMINTS website http://emints.org/wp-content/uploads/2014/09/The-Impact-of-eMINTS-Professional-Development-on-Teacher-Instruction-and-Student-Achievement-%E2%80%93-Year-1-Report-January-2013.pdf

*Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement, 75*(2), 311-337. doi: 10.1177/0013164414539163

*Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement, 73*(5), 757-783. doi:10.1177/0013164413486987

*Caspersen, J., & Raaen, F. D. (2014). Novice teachers and how they cope. *Teachers and Teaching, 20*(2), 189-211. doi: 10.1080/13540602.2013.848570

*Cherng, H. Y. S., & Halpin, P. F. (2016). The importance of minority teachers: Student perceptions of minority versus White teachers. *Educational Researcher, 45*(7), 407-420. doi: 10.3102/0013189X16671718

*Culp, K. M., Martin, W., Clements, M., & Presser, A. L. (2015). Testing the impact of a pre-

instructional digital game on middle-grade students' understanding of photosynthesis. *Technology, Knowledge and Learning, 20*(1), 5-26. doi: 10.1007/s10758-014-9233-5

Danielson, C. (2011). *Enhancing Professional Practice: A Framework for Teaching.* Alexandria, VA: ASCD.

*Donaldson, M. L., LeChasseur, K., & Mayer, A. (2017). Tracking instructional quality across secondary mathematics and English Language Arts classes. *Journal of Educational Change, 18*(2), 183-207. doi: 10.1007/s10833-015-9269-x

Darling-Hammond, L., & Rothman, R. (2015). Developing effective teaching: Why search globally? In L. Darling-Hammond, & R. Rothman (Eds.), *Teaching in the Flat World: Learning from High-Performing Systems* (pp. 1-5). New York: Teachers College Press.

Downer, J. T., Jamil, F., Maier, M. F., & Pianta, R. C. (2012). Implications of information processing theory for professional development of early educators. In C. Howes, B.K. Hamre, & R. C. Pianta (Eds.) *Effective Early Childhood Professional Development: Improving Teacher Practice and Child Outcomes.* Baltimore, MD: Paul H. Brookes.

*Durksen, T. L., Way, J., Bobis, J., Anderson, J., Skilling, K., & Martin, A. J. (2017). Motivation and engagement in mathematics: a qualitative framework for teacher-student interactions. *Mathematics Education Research Journal, 29*(2), 163-181. doi: 10.1007/s13394-017-0199-1

Ehrenberg, R., Brewer, D., Gamoran, A., & Willms, J. (2001). Does Class Size Matter? *Scientific American, 285*(5), 78-85. Retrieved from http://www.jstor.org.login.ezproxy.library.ualberta.ca/stable/26059424

Emmer, E., & Stough, L. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist, 36*, 103–112. doi:10.1207/S15326985EP3602_5

*Friedman, S. (2014). *Teacher emotional intelligence and the quality of interactions with students* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Global. Rutgers University, New Jersey.

*Gamlem, S. M., & Munthe, E. (2014). Mapping the quality of feedback to support students' learning in lower secondary classrooms. *Cambridge Journal of Education, 44*(1), 75-92. doi: 10.1080/0305764X.2013.855171

*Gill, B., Shoji, M., Coen, T., & Place, K. (2016). The content, predictive power, and potential bias in five widely used teacher observation instruments. *Mathematica Policy Research.* Retrieved from Regional Educational Laboratory website at http://ies.ed.gov/ncee/edlabs.

*Gitomer, D., Bell, C., Qi, Y., McCaffrey, D., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record, 116*(060304), 1-32.

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal, 26*, 91–108. doi: 10.1111/j.1471-1842.2009.00848.x

*Gregory, A., Allen, J. P., Mikami, A. Y., Hafen, C. A., & Pianta, R. C. (2014). Effects of a professional development program on behavioral engagement of students in middle and high school. *Psychology in the Schools, 51*(2), 143-163.

*Gregory, A., Allen, J. P., Mikami, A. Y., Hafen, C. A., & Pianta, R. C. (2015). The promise of a teacher professional development program in reducing racial disparity in classroom exclusionary discipline. In D. Losen (Ed.) *Closing the school discipline gap: Equitable remedies for excessive exclusion*, 166-179. New York, NY: Teachers College Press.

*Gregory, A., Hafen, C. A., Ruzek, E., Mikami, A. Y., Allen, J. P., & Pianta, R. C. (2016). Closing the racial discipline gap in classrooms by changing teacher practice. *School Psychology Review, 45*(2), 171-191.

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher, 43*(6), 293-303. doi: 10.3102/0013189X14544542

*Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English

language arts and teachers' value-added scores. *American Journal of Education, 119*(3), 445-470. Retrieved from http://www.jstor.org/stable/10.1086/669901

*Hafen, C. A., Allen, J. P., Mikami, A. Y., Gregory, A., Hamre, B., & Pianta, R. C. (2012). The pivotal role of adolescent autonomy in secondary school classrooms. *Journal of Youth and Adolescence, 41*(3), 245-255. doi: 10.1007/s10964-011-9739-2

*Hafen, C. A., Hamre, B. K., Allen, J. P., Bell, C. A., Gitomer, D. H., & Pianta, R. C. (2015). Teaching through interactions in secondary school classrooms: Revisiting the factor structure and practical application of the Classroom Assessment Scoring System–Secondary. *The Journal of Early Adolescence, 35*(5-6), 651-680. doi: 10.1177/0272431614537117

*Hafen, C. A., Ruzek, E. A., Gregory, A., Allen, J. P., & Mikami, A. Y. (2015). Focusing on teacher–student interactions eliminates the negative impact of students' disruptive behavior on teacher perceptions. *International Journal of Behavioral Development, 39*(5), 426-431. doi: 10.1177/0165025415579455

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials In Quantitative Methods For Psychology, 8*(1), 23. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3402032/

*Halpin, P. F., & Kieffer, M. J. (2015). Describing profiles of instructional practice: a new

approach to analyzing classroom observation data. *Educational Researcher, 44*(5), 263-277. doi: 10.3102/0013189X15590804

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development, 76*(5), 949-967. doi: 0009-3920/2005/7605-0001

Harms, T., & Clifford, R. (1998). *The Early Childhood Environment Rating Scale*. New York, NY: Teachers College Press.

Hattie, J. (2003). Teachers make a difference: What is the research evidence? *Distinguishing Expert Teachers from Novice and Experienced Teachers*. Camberwell: Australian Council for Educational Research. Retrieved from http://start.sd34.bc.ca/innovation/wp-content/uploads/2014/01/Hattie_TeachersMakeADifference.pdf

*Jamil, F. M., Downer, J. T., & Pianta, R. C. (2012). Association of pre-service teachers' performance, personality, and beliefs with teacher self-efficacy at program completion. *Teacher Education Quarterly, 39*(4), 119-138. Retrieved from http://files.eric.ed.gov/fulltext/EJ1001446.pdf

*Joe, J. N. McClellan, C. A., & Holtzman, S. L. (2014). Scoring design decisions: Reliability and the length and focus of classroom observations. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidelines from the Measures of*

Effective Teaching project (pp. 415-443). San Francisco: Josey-Bass. Retrieved from http://k12education.gatesfoundation.org/wp-content/uploads/2015/11/Designing-Teacher-Evaluation-Systems_freePDF.pdf

*Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Research Paper. MET Project. Bill & Melinda Gates Foundation. Retrieved from http://files.eric.ed.gov/fulltext/ED540960.pdf

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review, 47*, 180-195.

Kuncel, N. R., Credé, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. Review of Educational Research, 75(1), 63-82.

Lavigne, A. L., & Good, T. L. (2013). *Teacher and student evaluation: Moving beyond the failure of school reform.* New York: Routledge.

Luff, R., Byatt, D., & Martin, D. (2015). Review of the typology of research methods within the social sciences. National Centre for Research Methods Report. Retrieved from https://pdfs.semanticscholar.org/3c84/9e37fd4299eab0c5f1c0e98670f4b4ad99cf.pdf

*Malmberg, L.-E., & Hagger, H. (2009). Changes in student teachers' agency beliefs during a

teacher education year, and relationships with observed classroom quality, and day-to-day experiences. *British Journal of Educational Psychology, 79*, 677-694. doi: 10.1348/000709909X454814

*Malmberg, L.-E., Hagger, H., Burn, K., Mutton, T., & Colls, H. (2010). Observed classroom quality during teacher education and two years of professional practice. *Journal of Educational Psychology, 102*(4), 916-932. doi: 10.1037/a0020920

*Mashburn, A. J., Meyer, J. P., Allen, J. P., & Pianta, R. C. (2014). The effect of observation length and presentation order on the reliability and validity of an observational measure of teaching quality. *Educational and Psychological Measurement, 74*(3), 400-422. doi: 10.1177/0013164413515882

*McCaffrey, D. F., Yuan, K., Savitsky, T. D., Lockwood, J. R., & Edelen, M. O. (2015). Uncovering multivariate structure in classroom observations in the presence of rater errors. *Educational Measurement: Issues and Practice, 34*(2), 34-46. doi: 10.1111/emip.12061

*Meyer, C. (2011). *Decatur, IL Eisenhower High School CLASS-S observations.* Research report. Naperville, IL: American Institutes for Research.

*Meyers, C. V., Molefe, A., Brandt, W. C., Zhu, B., & Dhillon, S. (2016). Impact results of the eMINTS professional development validation study. *Educational Evaluation and Policy*

Analysis, 38(3), 455-476. doi: 10.3102/0162373716638446

*Mihaly, K., & McCaffrey, D. F. (2014). Grade-level variation in observational measures of teacher effectiveness. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidelines from the Measures of Effective Teaching project* (pp. 9-49). San Francisco: Josey-Bass. Retrieved from http://k12education.gatesfoundation.org/wp-content/uploads/2015/11/Designing-Teacher-Evaluation-Systems_freePDF.pdf

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). A composite estimator of effective teaching. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from https://www.nbexcellence.org/cms_files/resources/Jan%202013%20%20A%20Composite%20Estimator%20of%20Effective%20Teaching%20Research%20Paper.pdf

Moher D, Liberati A, Tetzlaff J, Altman DG, & The PRISMA Group. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA statement. *PLoS Medicine, 6*(7): e1000097. doi: https://doi.org/10.1371/journal.pmed.1000097

National Center for Teacher Effectiveness. (2012). *Mathematical Quality of Instruction.* Retrieved from http://isites.harvard.edu/icb/icb.do?keyword=mqi_training&tabgroupid=icb.tabgroup120173

*Neergaard, L., & Smith, T. (2012). *Comparing beginning teachers' instructional quality growth on subject-specific and global measures.* Proceedings from the Society for Research on Educational Effectiveness. Retrieved from http://files.eric.ed.gov/fulltext/ED535664.pdf

*Ozer, E. J., & Douglas, L. (2015). Assessing the key processes of youth-led participatory research: Psychometric analysis and application of an observational rating scale. *Youth & Society, 47*(1), 29-50. doi: 10.1177/0044118X12468011

*Park, Y. S., Chen, J., & Holtzman, S. L. (2014). Evaluating efforts to minimize rater bias in scoring classroom observations. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidelines from the Measures of Effective Teaching project* (pp. 383-414). San Francisco: Josey-Bass. Retrieved from http://k12education.gatesfoundation.org/wp-content/uploads/2015/11/Designing-Teacher-Evaluation-Systems_freePDF.pdf

Peters, M. D. J., Godfrey, C. M., Khalil, H., McInerney, P., Parker, D., & Soares, C. B. (2015). Guidance for conducting systematic scoping reviews. *International Journal Of Evidence-Based Healthcare, 13*(3), 141-146. doi: 10.1097/XEB.0000000000000050

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109-119.

Pianta, R. C., Hamre, B. K., & Mintz, S. (2012). *Classroom Assessment Scoring System: Secondary manual.* Charlottesville, VA: Teachstone.

Pianta, R. C., La Paro, K., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS).* Baltimore, MD: Paul H. Brookes.

*Ripski, M. B., LoCasale-Crouch, J., & Decker, L. (2011). Pre-service teachers: Dispositional traits, emotional states, and quality of teacher-student interactions. *Teacher Education Quarterly, 38*(2), 77-96. Retrieved from http://files.eric.ed.gov/fulltext/EJ926861.pdf

*Rutledge, S., Cohen-Vogel, L., & Osborne-Lampkin, L. (2012). *Identifying the characteristics of effective high schools: Report from year one of the National Center on Scaling Up Effective Schools.* Research report. National Center on Scaling Up Effective Schools (NJ1). Retrieved from http://files.eric.ed.gov/fulltext/ED538011.pdf

*Ruzek, E. A., Hafen, C. A., Allen, J. P., Gregory, A., Mikami, A. Y., & Pianta, R. C. (2016). How teacher emotional support motivates students: The mediating roles of perceived peer relatedness, autonomy support, and competence. *Learning and Instruction, 42*, 95-103. doi: 10.1016/j.learninstruc.2016.01.004

*Ruzek, E. A., Hafen, C. A., Hamre, B. K., & Pianta, R. C. (2014). Combining classroom observations and value added for the evaluation and professional development of teachers. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation*

systems: New guidelines from the Measures of Effective Teaching project (pp. 205-233). San Francisco: Josey-Bass. Retrieved from http://k12education.gatesfoundation.org/wp-content/uploads/2015/11/Designing-Teacher-Evaluation-Systems_freePDF.pdf

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68. doi: 10.1037110003-066X.55.1.68

Sanders, W. L., Wright, S. P., & Horn, S. P. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. Journal of Personnel Evaluation in Education, 11(1), 57-67. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.581.9316&rep=rep1&type=pdf

*Savitsky, T. D. & McCaffrey, D. F. (2014). Bayesian hierarchical multivariate formulation with factor analysis for nested ordinal data. *Psychometrika, 79*(2), 275-302. doi: 10.1007/S11336-013-9339-Z

Shyamalan, M. N. (2013). *I Got Schooled: The Unlikely Story of How a Moonlighting Movie Maker Learned the Five Keys to Closing America's Education Gap*. New York: Simon and Schuster.

*Shell, M. D., Gazelle, H., & Faldowski, R. A. (2014). Anxious solitude and the middle school transition: A diathesis x stress model of peer exclusion and victimization trajectories.

Developmental Psychology, 50(5), 1569-1583. doi: 10.1037/a0035528

*Smith, T. M., Cannata, M., & Haynes, K. T. (2016). Reconciling data from different sources: Practical realities of using mixed methods to identify effective high school practices. *Teachers College Record, 118*(7), 1-34.

*Smith, T., Preston, C., Haynes, K., & Neergaard Booker, L. (2015). Understanding differences in instructional quality between high and lower value-added schools in a large urban district. *Teachers College Record, 117*(11), 1-38.

*Spearman, J., & Watt, H. M. G. (2013). Perception shapes experience: The influence of actual and perceived classroom environment dimensions on girls' motivations for science. *Learning Environments Research, 16*, 217-238. doi: 10.1007/s10984-013-9129-7

UNESCO. (2014). *Teaching and Learning: Achieving Quality for All.* Retrieved from http://unesdoc.unesco.org/images/0022/002256/225654e.pdf

Vermette, D., Foote, D., Bird, C., Mesibov, D., Harris-Ewing, D., & Battaglia, D. (2001). Understanding constructivism(s): a primer for parents and school board members. *Education, 122*, 87–93.

*Virtanen, T. E., Lerkkanen, M.-K., Poikkeus, A.-M., & Kuorelahti, M. (2015). The relationship between classroom quality and students' engagement in secondary school. *Educational Psychology: An International Journal of Experimental Educational Psychology, 35*(8),

963-983. doi: 10.1080/01443410.2013.822961

*Wilson, H. K., Scult, M., Wilcher, M., Chudnofsky, R., Malloy, L., Drewel, E., ... & Denninger, J. W. (2015). Teacher-led relaxation response curriculum in an urban high school: impact on student behavioral health and classroom environment. *Advances in Mind-body Medicine, 29*(2), 6-14.

Wragg, T., & Wragg, E. C. (2012). *An Introduction to Classroom Observation.* New York, NY: Routledge.

Wubbels, T. (2014). Teacher-student relationships in class: A future for research [Abstract]. *Pedagogische Studien, 91*(5), 352-363.

*Yoder, N. P. (2013). *Linking social-emotional learning to teaching practices in an urban context: A mixed-methods study* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Global. (UMI 3610176).

*Yuan, K., McCaffrey, D. F., & Savitsky, T. D. (2013). *Analyzing the factorial structure of the Classroom Assessment Scoring System-Secondary using a Bayesian hierarchical multivariate ordinal model.* Paper presented at the Society for Research on Educational Effectiveness, Washington, DC. Retrieved from http://files.eric.ed.gov/fulltext/ED563074.pdf

Zohar, A., & Dori, Y. (2003). Higher order thinking skills and low achieving students: Are they

mutually exclusive? *Journal of the Learning Sciences, 12*, 145–181.

doi:10.1207/S15327809JLS1202_1

Chapter Three. Exploring the Preliminary Reliability and Validity of the Classroom Assessment Scoring System – Secondary (CLASS-S) in Junior High Schools in Alberta, Canada

Given the recognition that education must transform to keep up with the emerging demands of the 21st century, there has been an increasing emphasis on accountability and capacity building in education systems worldwide (e.g., Government of Alberta, 2010; Preston, Goldring, Guthrie, Ramsey, & Huff, 2017; Westergård, Ertesvåg, & Rafaelsen, 2018; World Education Forum, 2015). In particular, as researchers learn more about how teachers' classroom practices contribute significantly to student learning, education systems are striving to implement evidenced-based changes to support student achievement (e.g., Darling-Hammond & Rothman, 2015; Hamre & Pianta, 2005; Shyamalan, 2013). This means an increased emphasis on understanding the specific behaviours teachers engage in to facilitate learning in the classroom. One such way to deepen our understanding of teachers' classroom practices is through the use of standardized observational protocols as a means to capture and quantify classroom interactions.

Despite an extensive history of classroom observations in educational research, the last decade has seen an increased interest in the use of valid and reliable standardized classroom observational protocols (Pianta & Hamre, 2009). As a result, classroom observations have become increasingly more common and are becoming integral to accountability initiatives, and to inform capacity building, as a means to gather independent, systematic data (e.g., Allen,

Pianta, Gregory, Mikami, & Lun, 2011; Gregory, Hafen, Ruzek, Mikami, Allen, & Pianta, 2016; Kane & Staiger, 2012; Malmberg, Hgger, & Burn, 2010; Wragg & Wragg, 2012). One such systematic observational protocol, the Classroom Assessment Scoring System (CLASS: e.g., Pianta et al., 2008), has been widely adopted in both research and practice settings, to capture general pedagogical practices (Allen, Gregory, Mikami, Lun, Hamre, & Pianta, 2013).

The Classroom Assessment Scoring System

The CLASS is based on the Teaching Through Interactions Framework (TTIF), a model grounded in developmental and learning theories, and research (e.g., Hafen, Hamre, Allen, Bell, Gitomer, & Pianta, 2015; Pianta, 2006; Pianta & Hamre, 2009). The TTIF posits that teacher-student interactions in the classroom provide the foundation for learning and development, and that these interactions can be operationalized and observed. Furthermore, the CLASS is not subject specific, and, therefore, may contribute to our understanding of universal classroom practices. In order to reflect developmental theory and the fact that student needs evolve as they mature, the CLASS is available for various developmental stages beginning in infancy through to high school. In total, there are six versions of the CLASS, and although the general framework is similar across each age-level, behavioural indicators vary and are developmentally appropriate.

Much of the published research using the CLASS has focused on its use in early years

programs and elementary grades; however, the CLASS has increasingly been used in high school settings (see paper one). For example, the Classroom Assessment Scoring System – Secondary (CLASS-S; Pianta, Hamre, & Mintz, 2012) is specific to teacher-student interactions in Grades 7 through 12. To reflect adolescent development, the indicators of the CLASS-S place a greater emphasis on capturing various characteristics of adult and peer relationships, autonomy, and competence (e.g., Deci & Ryan, 2000; Hafen et al., 2015; Lavigne & Good, 2013). In the published research, the CLASS-S has been used in multiple ways, for example as an indicator of effective teaching (e.g., McCaffrey, Yuan, Savitsky, Lockwood, & Edelen, 2015), as a means to provide feedback to teachers (e.g., Smith, Hayes, Labonté, & Vargas, 2016), or as a model to guide interventions (e.g., Allen, Pianta, Gregory, Mikami, & Lun, 2011).

The CLASS-S is comprised of three domains associated with effective teaching: Emotional Support, Classroom Organization, and Instructional Support (Pianta & Hamre, 2009). Each domain includes three to five dimensions (see Table 3.1) and certified, trained observers rate various behavioural indicators of teacher-student interactions for each dimension (Pianta & Hamre, 2009). One dimension, Student Engagement, is not included in any of the domains as it solely emphasizes student behaviour and, therefore, does not capture a teacher-student

interaction (Pianta et al., 2012). Furthermore, Student Engagement is often considered an outcome of the CLASS-S domains (Bell et al., 2012).

In order to become a certified CLASS-S observer, coders must attend a two-day workshop and receive training from a master coder on how to rate each of the twelve CLASS-S dimensions using a 1-7 scale (Pianta et al., 2012). Every dimension includes specific behavioural indicators, and observers are trained to identify and score each dimension. Scores in the 1-2 point range are classified as low, in the 3-5 point range as mid, and in the 6-7 point range as high (Pianta et al., 2012). After training, observers must pass an online reliability test demonstrating they are able to code five video segments within one-point of a master coder, 80% of the time. CLASS-S certification is only valid for one year, and so coders must pass online reliability testing again in order to re-certify and continue using the CLASS-S protocol.

According to the developers, a standard observation is based on viewing a classroom lesson for approximately 15 to 20 minutes (either live or video-recorded), followed by time to assign the ratings for each dimension, before beginning another observation (Pianta et al., 2012). The 15 to 20 minute observation period is also referred to as a lesson segment, and so each lesson may include multiple segments depending on length of class time, and if the lesson was video-recorded, or observed live. Past research has identified strong correlations between a middle segment and the whole lesson score, suggesting that a segment from the middle of the

lesson may best represent the lesson as a whole (Joe, McClellan, & Holtzman, 2014). Other research has identified that two, 20-minute video-recorded segments coded in random order were found to be the most statistically sound measure of classroom practices (Mashburn, Meyer, Allen, & Pianta, 2014).

The CLASS-S in Research: Validity and Reliability

According to the Standards for Educational and Psychological Testing, validity "refers to the degree to which evidence and theory support the interpretations" of a measure (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014, p.11). There are various categories of evidence required to support the validity of the CLASS-S, and these include content validity, relationships with conceptually related constructs, relationships with criteria, and factor structure (AERA, APA, & NCME, 2014). In the manual, the developers have reported content-oriented evidence (see Pianta, Hamre, & Mintz, 2012); however, some researchers have questioned CLASS-S content. For example, Caspersen and Raaen (2014) were unable to corroborate CLASS-S content when they asked pre-service teachers their opinions of the relevance of the CLASS-S and if they were familiar with the various constructs from their teacher education programs. Further, Gamlem and Munthe (2014) reported on scoring criteria specific to the Quality of Feedback dimension, and proposed that instead of capturing learning-

oriented feedback, it emphasized encouragement-oriented feedback. In addition, limited evidence describing relationships between the CLASS-S and other related constructs has been identified. For example, significant correlations between similar observational protocols (e.g., Frameworks for Teaching; Gill, Shoji, Coen, & Place, 2016) have been noted. Also, researchers have found expected correlations between the Emotional Support domain and the Teacher Emotional Intelligence Measure (Friedman, 2014). Researchers have also identified significant correlations between the Classroom Organization domain and measures of classroom management efficacy beliefs, teaching-related stress, and teacher exhaustion (Virtanen, Pakarinen, Lerkkanen, Poikkeus, Siekkinen, & Nurmi, 2018). Finally, evidence is still needed to support the relationship between CLASS-S scores and predictions of performance, or the relationship with criteria. For example, Bell and colleagues (2012) concluded that evidence linking the CLASS-S to student achievement was divergent. Despite researchers identifying gains in student achievement after completing an intervention based on the CLASS-S framework (e.g., Allen et al., 2011; Allen, Hafen, Gregory, Mikami, & Pianta, 2015), and associations between increases in CLASS-S scores and value-added scores (e.g., Kane & Staiger, 2012), further research replicating and clarifying these relationships is required.

In contrast, explorations of the structural validity of the CLASS-S (i.e., factor structure) are more common in the literature. The CLASS-S was first used in published research in 2009

(see Malmberg & Haggar, 2009), at that time, the Emotional Support domain included Positive Climate, Teacher Sensitivity, Regard for Adolescent Perspectives, and Negative Climate. The Classroom Organization domain included Behaviour Management, Productivity, and Instructional Learning Formats. The Instructional Support domain included Content Understanding, Analysis and Problem Solving (later renamed Analysis and Inquiry), Quality of Feedback, and Procedures and Skills. Notably, the CLASS-S structure varied over subsequent publications (see paper one), resulting in the manualized version (Table 3.1) published in 2012 (Pianta et al., 2012). The final version moved Negative Climate to the Classroom Organization domain, and Instructional Learning Formats to the Instructional Support domain. Furthermore, the Instructional Dialogue dimension was newly added to the protocol under the Instructional Support domain. As a result of ongoing structural changes in the early years of its use, it can be difficult to make definitive comparisons across studies using the CLASS-S.

The published version of the CLASS-S was based on a factor analysis completed by Hafen and colleagues (2015), supporting the theoretical three domain factor structure, and finalizing the assignment of dimensions to each factor as seen in the manual (see Table 3.1). Subsequently, other researchers have explored the CLASS-S factor structure and identified modified versions of the three-factor model as better fitting their respective data sets (e.g., in the United Kingdom; Malmberg, Hagger, Burn, Mutton, & Colls, 2010; in Finland; Virtanen et al.,

2018). Furthermore, some researchers have challenged the three-factor model and have preferred a two-factor (in the United States; McCaffrey, Yuan, Savitsky, Lockwood, & Edelen, 2015; Yuan, McCaffrey, & Savitsky, 2013), or proposed a single-factor model as a better indicator of teaching effectiveness (in the United States; Kane et al., 2012; McCaffrey et al., 2015; and in the United Kingdom; Malmberg, et al., 2010;). Researchers have also identified different factor structures at different rating levels, for example at the individual segment level score, in contrast to an aggregated, overall teacher score (McCaffrey et al., 2015; Yuan, McCaffrey, & Savitsky, 2013). Researchers have argued averaging scores across multiple data points discounts variability in scores, and the inherent nested nature of the data (Huang, 2017; McCaffrey et al., 2015; Westergård et al., 2018). Therefore, other models including bifactor (Hafen et al., 2015), and multilevel factor analyses (e.g., Malmberg et al., 2010; Westergård et al., 2018) have been explored in an effort to account for the complexities (e.g., nested data, sources of error) associated with measuring teacher-student interactions in the classroom.

Table 3.2 provides a brief summary of all studies identified that use the domains and dimensions of the published CLASS-S (with the exception of Student Engagement). It should be noted that these studies represent a small sample of the total number of publications using the CLASS-S (see paper one) that also included descriptive statistics (i.e., means) for the domains and/or dimensions. The studies using the published CLASS-S included three published articles

(one from the United States, one from Finland, and one from Norway), one unpublished study (from Canada), and one dissertation (from the United States). Although comparisons across the study means are not statistically possible, patterns in scores appear to be similar, with higher scores reported for Emotional Support (and associated dimensions), and lower scores reported for Instructional Practices (and associated dimensions) identified cross-culturally.

Despite the challenge making comparisons across versions of the CLASS-S, some researchers have provided emerging evidence to support the reliability of the CLASS-S (e.g., Hafen et al., 2015). According to the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014), reliability includes both reliability/precision (i.e., how consistent scores are across time and procedures), and reliability coefficients (i.e., statistical measures of correlation between test scores). Researchers have explored the reliability of the CLASS-S by examining item and score reliabilities, and intra-rater reliability.

In an effort to report reliability coefficients, many researchers have included item, scale, and inter-rater reliabilities (IRRs). Notably, researchers have consistently reported acceptable item and scale reliabilities (e.g., Allen et al., 2013; Gregory, Allen, Mikami, Hafen, & Pianta, 2014; Virtanen et al., 2018), while procedures exploring inter-rater reliabilities have varied.

Inter-rater reliability results from a common procedure many researchers used, whereby they double-coded all or a percentage of classroom segments to explore consistency across coders. Statistical procedures to calculate IRR included: intra-class correlation coefficients (ICCs; e.g., Allen et al., 2013; Bell et al., 2012; Malmberg et al., 2010; Virtanen et al., 2018), percentage of agreement within 1 point (e.g., Booker, 2014; Gregory et al., 2016), and Cohen's kappa coefficient (e.g., Hafen et al., 2015). Researchers consistently reported IRR statistics within acceptable ranges, for example ICCs > 0.40 (Cicchetti, 1994), and agreement within 1 point >80% (Pianta et al., 2012).

Education in Alberta, Canada

In the most recent international comparison of student performance outcomes, the Program for International Student Assessment (PISA; The Organization for Economic Co-Operation and Development [OECD], 2015) ranked Canada among the top educational systems in the world. However, in Canada, education is primarily governed by each province or territory autonomously and, therefore, varies across the country (Coughlan, 2017). As a consequence, if Canadian provinces submitted their PISA scores as separate entries to OECD, three separate provinces (Alberta, British Columbia, and Quebec) would have ranked among the top five worldwide entries for student performance in science (Coughlan, 2017). This means Canada's education system is strong, and is receiving international recognition.

In Canada, and in Alberta in particular, the general educational context is similar to that of the United States (U.S.; e.g., overall levels of government funding, class size, grades/ages; OECD, 2015). Alberta schools fall under either the public or separate school boards and both boards follow provincial educational standards. The main difference is that the separate school board includes the integration of faith-based curricula. Provincial Achievement Testing is administered to all students in grades 6 and 9 to monitor student learning, and scores are meant to reflect a provincial standard of achievement across all school jurisdictions (Alberta Education, 2018).

In 2010, the Government of Alberta recognized that despite having a world-class education system, there was a need to transform education to meet future needs (Government of Alberta, 2010). With an emphasis on 21st century learning skills, the Government of Alberta outlined principles for education that included learning-centered, accountable, engaged, inclusive, responsive, sustained, and innovative approaches (Government of Alberta, 2010). As a means to support the Government of Alberta's new direction, the Learning and Technology Policy Framework was subsequently published (Alberta Education, 2013). The Learning and Technology Policy Framework (2013) highlighted five policy directions for the province:

student-centered learning, research and innovation, professional learning, leadership, and access, infrastructure and digital learning environments.

The present study. As part of the Government of Alberta's transformation of the education system, the province sought to fund the implementation of technology-based innovations in junior high classrooms across the province. One such technology-based project was *Flexible Pathways to Success: Technology to Design for Diversity* (Flexible Pathways; Smith et al., 2016), a two-year collaborative project that included Alberta Education, five unique school jurisdictions, and researchers from the University of Alberta. The aim of the Flexible Pathways project was to support the implementation of technology to meet the diverse learning needs in inclusive classroom settings. The role of the researchers was to support the implementation of the initiatives, by taking a developmental evaluation approach (Patton, 2011) in an effort to gather data and feedback results on an ongoing basis to inform both formative and summative findings at the administrative, classroom, and individual student levels. Using mixed-methods methodology, the researchers gathered a wide-range of data (e.g., interviews, focus groups, questionnaires, video recorded classroom lessons) from a variety of sources (e.g., administrators, teachers, students). The research framework was informed by several theories in education, including the Substitution, Modification, and Redefinition Model (e.g., Puentadura,

2010), Universal Design for Learning (e.g., Rose & Meyer, 2006), and the Teaching Through Interactions Framework (e.g., Hafen et al., 2015).

The present study focuses on the video recorded classroom lessons to explore the preliminary validity of the Teaching Through Interactions Framework (i.e., based on the CLASS-S protocol) in a sample of junior high school classrooms from various school jurisdictions across Alberta, Canada. As such, this study contributes to the emerging body of literature examining the use of the CLASS-S in a variety of cross-cultural contexts outside of the United States (e.g., Finland; Virtanen et al., 2018; Norway; Westergård et al., 2018). This study will therefore report on the reliability (i.e., the item and scale reliabilities, and IRR of the CLASS-S), and structural validity (i.e., the factor structure of the CLASS-S), in a diverse Canadian context.

Method

Participants

Participants included 38 Grade 7 and 8 classroom teachers (71% female) from nine schools, representing five diverse school jurisdictions across Alberta, Canada. Mean age for participating teachers was 34.48 (standard deviation [SD] = 8.37, range from 24 to 61 years of age), with a mean teaching experience reported of 8.32 years (SD = 7.84, range from 1 to 36 years). All teachers had a minimum of a Bachelor's degree, with three teachers indicating they

had also obtained a Master's degree. All teachers had volunteered to participate in the larger Flexible Pathways project, and consented to video recordings of their classroom instruction. Subjects taught included language arts, science, social studies, mathematics, and one theoretical physical education lesson. As previously noted, the Flexible Pathways project was a two-year initiative, and therefore natural attrition occurred as some teachers moved to different schools or grades, and some projects chose to engage new teachers in the second year as a means to grow their projects. Therefore, a total of 26 teachers participated in year one, with 16 teachers continuing their participation into year two. In addition, 12 new teachers joined the project in year two.

Students were also included in video recorded lessons as long as their parents had provided consent as part of the larger project, and if the student assented to participate. Students who did not have consent to be video recorded sat outside of the visual range of the video-camera so that they would not miss any instructional time. In total, 618 students consented to participate in the larger Flexible Pathways project.

Measures

Teaching Through Interactions Framework: The CLASS-S. The CLASS-S (Pianta et al., 2012) is an observational protocol coders use to assess classroom interactions, as outlined in the Teaching Through Interactions Framework (TTIF). As noted previously, the TTIF

categorizes classroom interactions into three broad domains: Emotional Support, Classroom Organization, and Instructional Support. According to the factor structure reported in the manual, each domain is comprised of between three and five specific dimensions. Further, the manual describes a number of behavioral indicators for each dimension, and provides specific guidelines on how to score dimensions along a one to seven scale (1-2 = low range, 3-5 = mid range, and 6-7 = high range). In addition, the CLASS-S protocol includes a dimension of Student Engagement, which has been commonly considered an outcome of the three broad domains and, therefore, has not been included in past studies exploring the overall factor structure (e.g., Hafen et al., 2015; McCaffrey et al., 2015; Virtanen et al., 2018; Westergård et al., 2018). The Student Engagement dimension was included for descriptive purposes in the present study, but was not included in factor analyses.

Coder training. Coders included the principal investigator of the Flexible Pathways project and a team of seven masters and doctoral research assistants. All coders attended a two-day CLASS-S training workshop conducted by a master trainer. Following training, coders were required to complete an online reliability test requiring them to code five videos within one-point of the master codes, 80% of the time. Only those who passed the online reliability test to qualify as a Certified Secondary CLASS observer were eligible to code videos. In an attempt to minimize scoring drift over the course of the project, coders engaged in calibration exercises

prior to scoring year two videos (i.e., fall 2015). Coders were required to score two videos that had been double-coded in year one with exceptional interrater reliability (i.e., within one point, >80% of the time). Feedback and additional coaching were provided in the event coders were unable to code within one point, at least 80% of the time. All coders met acceptable coding standards prior to completing any video coding for the project.

Procedure. The researchers worked with school administration to facilitate classroom visits to video record at least two complete classroom sessions (between 45 to 60 minutes depending on the school's structure) for each participating teacher. Teachers were given notice prior to any videotaping, and were asked to ensure a typical instructional session was planned (i.e., not test taking). Teachers wore a lapel microphone to better capture teacher-student interactions. Classroom observations were video recorded at four time points across the two-year Flexible Pathways project (i.e., fall 2013, spring 2014, fall 2014, and spring 2015). Video recordings were divided into 15 to 20 minute segments, as recommended in the CLASS-S manual (Pianta, Hamre, & Mintz, 2012) and 3 segments per classroom were scored, per time point, for a total of 291 segments.

Data Analysis

Data analyses (e.g., descriptive statistics, reliability coefficients, exploratory factor analyses) were conducted using IBM SPSS Statistics version 25 (SPSS). Confirmatory factor

analyses and multilevel confirmatory factor analyses were conducted using lavaan version 0.5-23.1097 (Rosseel, 2017) in R version 3.5.0 (R Core Team, 2018) and RStudio version 1.1.442 (RStudio Team, 2018).

Inter-rater reliability. Twenty percent of the video segments were randomly selected to be double coded by two independent coders (i.e., 59 video segments) in order to calculate inter-rater agreement (IRA). According to the CLASS-S manual, IRA is considered acceptable if scores between the coders fall within one point (i.e., +/-1), 80% of the time (Pianta et al., 2012). However, according to Hallgren (2012), IRA is not a sufficient measure of inter-rater reliability (IRR). Therefore, based on Hallgren's (2012) recommendations, the intra-class correlation coefficient (ICC) was also calculated as an indicator of IRR. A two-way, random effects model was defined according to the study design (Hallgren, 2012). ICCs are considered to be an excellent rating of agreement when the value falls between 0.75 and 1.0, good when it falls between 0.60 and 0.74, fair when it falls between 0.40 and 0.59, and poor if it falls below 0.40 (Cicchetti, 1994).

Factor analysis. The factor structure was investigated using two different approaches, the first examined the factors at the classroom level, whereby CLASS-S ratings were averaged for each classroom teacher across segments, and time points to obtain an overall aggregated dimension score. This method is considered to be the more traditional approach, but has drawn

criticism due to the nested nature of the data (e.g., segments nested within lessons, within classrooms; Westergård et al., 2018). The second approach examined the factors at the rating level, whereby all available ratings were included in the analysis. Importantly, researchers have identified averaging scores at the teacher level may lead to a different factor structure (e.g., a one-factor model) than at the rating level (e.g., a two-factor model; McCaffrey, Yuan, Savitsky, Lockwood, & Edelen, 2015; Yuan, McCaffrey, & Savitsky, 2013).

Furthermore, structural validity was examined through various factor analyses, including more complex multilevel models (e.g., within-level segment ratings, and between-level teacher scores). The process for data analysis included various steps. First, two exploratory factor analyses (EFA) were conducted, one at the classroom level (averaged scores across dimensions for each teacher) and one at the rating level. Maximum likelihood estimation was applied according to best practice guidelines (Costello & Osborne, 2005). Next, confirmatory factor analyses were conducted to assess various factor structures (i.e., model identified in the EFA conducted here, and factor structures reported in past researcher), at both the teacher and rating levels. Finally, multilevel confirmatory factor analyses were explored (Huang, 2017).

The same criteria for goodness-of-fit identified in previous CLASS-S factor studies (i.e., Hafen et al., 2015; Westergård et al., 2018) was applied to the current analyses and included the following indicators: comparative fit index (CFI) above 0.90 and close to 0.95, Tucker-Lewis

Index (TLI) above 0.90 and close to 0.95, standardized root mean square residual (SRMR) under 0.12 and close to 0.08, and root mean square error of approximation (RMSEA) under 0.10 and close to 0.06 (Hu & Bentler, 1999; Schermelleh-Engel, Moosbrugger, & Müller, 2003). The chi-square test (X^2) was also included, although traditionally a non-significant X^2 indicated an adequate model, researchers have identified this measure as an inadequate indicator of model fit due to sensitivity to sample size, and model complexity (Schermelleh-Engel et al., 2003).

Results

Descriptive statistics for all domains (based on the CLASS-S manual; Pianta et al., 2012) and dimensions are presented in Table 3.3. Prior to further analysis, all variables were assessed for normality, no missing data or outliers were identified, and kurtosis or skewness fell within acceptable values (e.g., +/-1.96; Field, 2016).

Reliability

Item and scale reliabilities. Correlations between CLASS-S items are provided in Table 3.4. Cronbach's alpha estimates for the original three-factor structure scales were 0.76 for Emotional Support, 0.69 for Classroom Organization, and 0.86 for Instructional Support, reflecting acceptable item reliabilities (i.e., > 0.65; Vaske, Beaman, & Sponarski, 2017). Cronbach's alpha estimates for the alternative model (see below) remained consistent with the

original factor structure (i.e., 0.76 for Emotional Support, 0.70 for Classroom Organization, and 0.86 for Instructional Support).

Inter-rater reliability. IRR was calculated using two methods, percent agreement between coders within one-point (PWO; Pianta et al., 2012), and ICCs (Hallgren, 2012). Table 3.5 summarizes the IRR results. Overall, the percent agreement between two coders that fell within one-point (82.2%) was acceptable according to the CLASS-S developers (Pianta et al., 2012). The most challenging dimensions for coders to score reliably within one-point were those that comprised the Instructional Support domain (i.e., Content Understanding, Analysis and Inquiry, Quality of Feedback, and Instructional Dialogue) and the Productivity dimension. Based on the ICCs, the majority of the dimensions fell in the fair to good range for inter-rater consistency (i.e., 0.40 – 0.74; Cicchetti, 1994), with one dimension (Regard for Adolescent Perspective) falling in the excellent range and two dimensions (Productivity, Analysis and Inquiry) falling in the poor range.

Structural Validity

Exploratory factor analyses. An EFA using maximum likelihood extraction was completed using all data (i.e., at the rating level) and suggested a three-factor model that accounted for 68% of the total variance. The three factors were similar to those identified in the manual (i.e., Emotional Support, Classroom Organization, Instructional Support), with some

minor changes. The EFA completed on the current sample indicated that Regard for Adolescent Perspective loaded equally on both the Emotional Support and Instructional Support domains, while the Negative Climate was more strongly associated with the Emotional Support domain. Therefore, Regard for Adolescent Perspective was omitted, and Negative Climate was loaded on the Emotional Support domain in the alternative model at the rating level.

A second EFA using maximum likelihood extraction with varimax rotation was conducted at the classroom level, using the averaged scores from each teacher. Results of the EFA suggested a two-factor model, accounting for 69% of the total variance. The two factors were similar to a modified structure previously reported by Yuan, McCaffrey, and Savitsky (2013), where Classroom Organization items (i.e., Behavior Management, Productivity, Negative Climate) remained loaded on one factor, and all other items loaded on a second factor they called Teaching Capacity.

Confirmatory factor analyses and multilevel confirmatory factor analyses. Various CFAs were performed to identify which factor structure(s) from the EFAs, and those reported in previous studies, best fit the current Canadian sample. Table 3.6 provides the model fit indices for the confirmatory factor analyses performed. At both the rating level and classroom level, the original three-factor model (i.e., Hafen et al., 2015; Pianta et al., 2012) provided a poor fit to the data. Furthermore, a one-factor model (e.g., McCaffrey et al., 2015; Kane & Staiger, 2012) was

also a poor fit at both the rating and classroom levels. Additionally, a two-factor model based on the EFA previously identified at the classroom level provided a poor fit the present data.

At the rating level, two three-factor models provided an acceptable fit (see Table 3.6). A modified model based on the factor structure identified in a sample of secondary schools in Finland (see Virtanen et al., 2018) was similar to the published CLASS-S factor structure with the exception that the Regard for Adolescent Perspective, and Instructional Learning Formats dimensions were excluded. The model was improved based on the modification indices, allowing for a residual correlation between the Negative Climate and Positive Climate dimensions. This modification was also made to improve the fit in the original articles (Virtanen et al., 2018). The modified three-factor model is represented in Figure 3.1. The fit of the modified three-factor model was: $X^2(23) = 77.73$, CFI = 0.95, TLI = 0.92, RMSEA = 0.09, SRMR = 0.05. Factor loadings ranged from 0.57 to 0.86. Correlations between the domains based on the modified model were 0.53 between Emotional Support and Classroom Organization, 0.34 between Classroom Organization and Instructional Support, and 0.44 between Emotional Support and Instructional Support.

An alternative three-factor based on the previously described EFA, also provided an acceptable fit to the present data at the rating level. The alternative three-factor model is represented in Figure 3.2. Notably, this model was also similar to the published CLASS-S factor

structure with the exception that the Regard for Adolescent Perspective dimension was excluded, and the Negative Climate dimension was moved to the Emotional Support domain. The model was improved based on modification indices and theory, by allowing the residuals of the Negative Climate and Teacher Sensitivity dimensions, and the Negative Climate and Behavior Management dimensions to correlate. The fit of the alternate three-factor model was: $X^2(30) = 101.56$, CFI = 0.94, TLI = 0.91, RMSEA = 0.09, SRMR = 0.05. Factor loadings ranged from 0.60 to 0.89. Correlations between the domains based on the alternate model were 0.50 between Emotional Support and Classroom Organization, 0.34 between Classroom Organization and Instructional Support, and 0.45 between Emotional Support and Instructional Support.

At the classroom level, one model based on the modified three-factor model previously described, also provided an acceptable fit to the present Canadian sample. The modified three-factor model is represented in Figure 3.3. The fit of the modified three-factor model was: $X^2(23) = 35.98$, CFI = 0.94, TLI = 0.91, RMSEA = 0.12, SRMR = 0.08. Although it should be noted that RMSEA was slightly above the recommended cutoff, likely as a result of the small sample size at the classroom level, (Hu & Bentler, 1999; Schermelleh-Engel, Moosbrugger, & Müller, 2003). Factor loadings ranged from 0.70 to 0.99. Correlations between the domains based on the modified model were 0.51 between Emotional Support and Classroom Organization, 0.48

between Classroom Organization and Instructional Support, and 0.65 between Emotional Support and Instructional Support.

Due to the nested nature of the present data, a multilevel confirmatory factor analysis (MCFA) was warranted and recommendations based on Huang (2017) were followed. The present data was not a large enough sample at the classroom level to adequately execute an MCFA. Furthermore, the sample was not large enough to explore a bi-factor model.

Discussion

In summary, the present article reports on the reliability (i.e., the item and scale reliabilities, and IRR of the CLASS-S), and structural validity (i.e., the factor structure of the CLASS-S), in a sample of classrooms from the province of Alberta, a diverse Canadian context. Specifically, Cronbach's alpha's reported for the original three-factor scales were 0.76 for Emotional Support, 0.69 for Classroom Organization, and 0.86 for Instructional Support. Findings from this study and were similar to those reported in other studies. For example, Gitomer and colleagues (2014) reported Cronbach alpha's of 0.83 for Emotional Support, 0.75 for Classroom Organization, and 0.86 for Instructional Support in a sample from the United States. Furthermore, Virtanen and colleagues (2015) reported Cronbach's alphas of 0.86 for Emotional Support, 0.72 for Classroom Organization, and 0.91 for Instructional Support in their sample from Finland. Therefore, item and scale reliabilities have consistently reflected

acceptable Cronbach's alpha estimates (i.e., > 0.65; Vaske, Beaman, & Sponarski, 2017) across samples.

Two methods were used to calculate IRR in the present data, percent agreement between coders within one-point (Pianta et al., 2012), and ICCs (Hallgren, 2012). Both methods of IRR were consistent with past research findings. Overall, 20% of segment videos were double coded in the current sample, and percent agreement between coders that fell within one-point was 82.2%. According to CLASS-S developers, a minimum of 80% agreement is considered acceptable reliability (Pianta et al., 2012). Our findings were consistent with other studies, for example, Casabianca and colleagues (2015) double coded 20% of their segment videos and reported 82% agreement within one-point. In another study, Gregory and colleagues (2016) double coded all segment videos and reported 80.3% agreement within one-point.

The second method used to calculate IRR was ICCs (Hallgren, 2012). Based on ICCs, the majority of dimensions fell in the fair to good range for inter-rater consistency (i.e., 0.40 – 0.74; Cicchetti, 1994), with one dimension, Regard for Adolescent Perspective, falling in the excellent range (0.79), and two dimensions, Productivity (0.32), and Analysis and Inquiry (0.32) falling in the poor range. In comparison, past findings highlight variability in ICCs. For example, Bell and colleagues (2014) reported ICCs from two samples collected in the United States, the Measures of Effective Teaching (MET) project, and the Understanding Teaching

Quality (UTQ) project. ICCs were consistently lower in the UTQ project, with every dimension falling in the poor range (i.e., < 0.30), while ICCs for every dimension in the MET project fell in the fair to good range for inter-rater consistency (i.e., from 0.41 to 0.66). In a sample from Finland, researchers reported most ICCs were greater than 0.60, with two exceptions: Teacher Sensitivity (0.25), and Negative Climate (0.26; Virtanen et al., 2018). Further, in a sample from the United Kingdom, researchers reported all dimension ICCs fell in the fair to excellent range (i.e., ICCs ranged from 0.49 to 0.75). With few exceptions (e.g., the UTQ project), the ICCs for the majority of CLASS-S dimensions appear to fall within acceptable ranges (Cicchetti, 1994) across studies. However, there are some occasional poor ICCs reported for select dimensions. Notably, poor dimension reliabilities do not appear to be consistently identified across studies (i.e., no apparent pattern of poor dimension ICCs), suggesting there may be other factors impacting ICCs. For example, individual differences between coders (i.e., rater drift, error) may be associated with dimensional ICC variability (Casabianca et al., 2015). One pattern suggested by researchers is that dimensions associated with the Instructional Support domain may be the most difficult to code reliably (Bell et al., 2014; Gitomer et al., 2014), as they may rely more heavily on the coder's judgment and therefore be more susceptible to rater error and/or drift.

Results exploring and confirming the factor structure of the CLASS-S in the present data collected in Alberta, Canada, supported the theoretical three-factor structure of the CLASS-S

(i.e., Emotional Support, Organizational Support, and Instructional Support) with some small changes to the original factor structure proposed by the developers (Pianta et al., 2012; Hafen et al., 2015). The factor structure was examined at two levels, the segment level, and the aggregated classroom level, as researchers have argued averaging scores across multiple data points discounts variability in scores, and the inherent nested nature of the data (Huang, 2017; McCaffrey et al., 2015; Westergård et al., 2018). However, the current sample was not large enough to explore the fit of a multilevel or bi-factor model.

At the rating level, two three-factor models provided an acceptable fit (see Table 3.6), a modified model identified in a sample of secondary schools in Finland (Virtanen et al., 2018), and an alternative model identified in an EFA conducted in the present sample. Both the modified and alternative models were similar to the published CLASS-S factor structure with some exceptions. The modified model (based on Virtanen et al., 2018) excluded both the Regard for Adolescent Perspective and Instructional Learning Formats dimensions, and allowed a residual correlation between the Negative Climate and Positive Climate dimensions. The alternative three-factor (based on the results of the EFA), also provided acceptable fit to the present data at the rating level. Notably, this model was similar to the published CLASS-S factor structure with the exception that the Regard for Adolescent Perspective dimension was excluded, the Negative Climate dimension moved to the Emotional Support domain, and residuals for the

Negative Climate and Teacher Sensitivity dimensions, and the Negative Climate and Behavior Management dimensions were allowed to correlate. The fits for both identified models were comparable and goodness-of-fit statistics reported were acceptable (Hu & Bentler, 1999; Schermelleh-Engel, Moosbrugger, & Müller, 2003). However, the alternative three-factor model based on the EFA reported herein, might be considered a slightly stronger model as it excluded one dimension and not two as did the modified model (Virtanen et al., 2018). Notably, both models include one domain with only two dimensions loaded; most researchers recommend a minimum of three dimensions per domain for a stronger factor structure (Costello & Osborne, 2005). Despite differences at the dimension level, a three-factor model including the Emotional Support, Classroom Organization, and Instructional Support domains, has received the most support across international samples (e.g., United States; Hafen et al., 2015; United Kingdom; Malmberg et al., 2010; Finland; Virtanen et al., 2018; Norway; Westergård et al., 2018).

One modification both models made was to exclude the Regard for Adolescent Perspectives dimension. In both the present sample, and the study conducted by Virtanen and colleagues (2017), the Regard for Adolescent Perspectives dimension cross-loaded on both the Emotional Support and Classroom Organizational domains. Others have proposed this dimension reflects both the provision of support for adolescent needs (e.g., emotional support, autonomy), and classroom practices (e.g., allowing choice, providing opportunities for

leadership; Virtanen et al., 2018). Another reason researchers proposed Regard for Adolescent Perspective may have had lower discriminant validity was that the Finish sample was based on Grade 6 classrooms, and at this grade level teachers may have been more directive in their teaching practices, and placed less of an emphasis on adolescent developmental needs (Virtanen et al., 2018). The present Canadian sample included both Grade 7 and 8 classrooms; however, it may also follow that many teachers placed a greater emphasis on classroom practices (e.g., allowing choice, providing opportunities for group or independent work), versus making real-world connections or developing supportive peer relationships.

Another modification made to the alternative model in the present sample at the rating level included moving the Negative Climate dimension to the Emotional Support domain. Interestingly, Negative Climate was originally associated with the Emotional Support domain in early CLASS-S versions (e.g., Hafen et al., 2015). It was moved to the Classroom Organization domain for the published version following factor analysis (Hafen et al., 2015; Pianta et al., 2012). In the current data, Negative Climate and Positive Climate were allowed to correlate; the same modification was made in other published research (e.g., Virtanen et al., 2018). As such, it is possible that Negative Climate and Positive Climate are more closely related, and may represent a single dimension. An additional modification was made in the current sample, where the residuals of the Negative Climate and Behaviour Management dimensions were also allowed

to correlate. Conceptually there may be overlap between these two dimensions, as one includes ratings of the use of punitive measures to control negative classroom behaviour (i.e., Negative Climate), and the other includes ratings of proactive and reactive classroom management strategies (i.e., Behaviour Management). Therefore, if Negative Climate was low, positive indicators of Behaviour Management were likely present. In fact, high correlations between Negative Climate (reverse scored) and Positive Climate (0.53, $p < 0.01$), and Negative Climate (reverse scored) and Behaviour Management (0.46, $p < 0.01$) were identified.

At the classroom level, results of the EFA identified a two-factor model that was similar to previous research proposing two domains: Classroom Organization (similar to original model), and Teaching Capacity (which included all other dimensions; Yuan, McCaffrey, & Savitsky, 2013). However, results of the CFA goodness-of-fit indices did not support this model in the present data. Furthermore, only one model, based on the modified three-factor model already described (Virtanen et al., 2018), provided an adequate fit to the present Canadian sample. Despite the small sample size at the classroom level, the original three-factor model was still the best fit when modifications suggested by Virtanen and colleagues (2017) were applied.

Limitations

Although the present sample included data from a diverse geographical area across Alberta, Canada, it was limited in size and therefore may not adequately represent the sum of the

Albertan educational context. As part of a larger project, the present sample included classrooms from jurisdictions engaged in implementing a variety of technology initiatives. Although the larger project occurred over two years, teachers' involvement varied across jurisdictions. For example, some teachers participated for the duration of the project, while others participated in either year one or year two. Thus, the number of observations conducted per classroom ranged from 3 segments at one time point (e.g., Fall 2013), to 3 segments across four time points (e.g., Fall 2013, Spring 2014, Fall 2014, and Spring 2015). Therefore, there may be differences between teachers who participated in different portions of the study. However, due to the limited sample size, all segments were included in this study for analysis. In addition, the diversity of the sample itself may have impacted the findings presented here, as such, future large-scale studies may allow for clearer comparisons across geographical areas. Furthermore, limitations related to sample size meant that more complex models, suggested to better represent the nested nature of the data, were not possible (e.g., multilevel factor analysis; Westergård et al., 2018).

Despite reporting preliminary support for the reliability and structural validity of the CLASS-S in the present context, other evidence related to content validity, relationships with conceptually related constructs, and relationships with criteria (AERA, APA, & NCME, 2014) were not explored. Therefore, further studies of the CLASS-S in Alberta should explore concurrent associations with other measures of effective classroom practices, and related

constructs (e.g., adolescent autonomy, teacher emotional intelligence, instructional practices). Further exploration of relationships between the CLASS-S and criteria, or predictive validity, would also be beneficial. For example, future studies could examine student outcomes, or the ability of the CLASS-S to predict positive student gains.

Conclusion

The reliability and structural validity of the CLASS-S has been explored in the United States (e.g., Hafen et al., 2015; Kane & Staiger, 2012), Finland (Virtanen et al., 2018), and Norway (Westergård et al., 2018). However, the current study represents the first to report on the preliminary reliability (i.e., item and scale reliabilities, and IRR) and structural validity (i.e., factor structure) in the context of a varied sample from Alberta, Canada. Thus, this paper contributes to the emerging body of literature supporting the use of the CLASS-S as a valid and reliable observational protocol for teacher-student classroom interactions. Reliability and validity were consistent with past findings, and suggest the CLASS-S is a valid tool for use across educational contexts in Alberta, Canada. Structural validity of the theoretical model was supported, with some modifications. Present findings were similar to those identified in a sample of classrooms from Finland. Finland has routinely been identified as a leader in education, and Canada has also been ranked by the OECD (2015) among the top educational systems in the world. Although further reliability and validity studies are needed to generalize

the use of the CLASS-S across Canada, the CLASS-S is likely a valuable tool to support Canadian educational systems as they strive to become leaders in educational transformation.

Table 3.1.

Domains, dimensions and descriptions of the Classroom Assessment Scoring System – Secondary (Pianta, Hamre, & Mintz, 2012).

Domains	Dimensions	Description
Emotional Support	Positive Climate	Degree of positive relationship (e.g., emotional connection, warmth, respect, shared enjoyment) between teacher and students
	Teacher Sensitivity	Degree of teacher's responsiveness to student and classroom needs (e.g., academic, social-emotional, developmental)
	Regard for Adolescent Perspective	Degree of teacher's ability to support and capitalize on adolescent needs and worldviews (e.g., support autonomy, provide opportunities for leadership, connections to real-world content)
Classroom Organization	Behaviour Management	Degree of teacher's ability to proactively and reactively manage classroom expectations and misbehaviour
	Productivity	Degree of teacher's ability to manage classroom time and routines
	Negative Climate	Degree of negativity (e.g., punitive control, disrespect) between teacher and students
Instructional Support	Instructional Learning Formats	Degree of teacher's use of materials, active facilitation, and provision of learning objectives to maintain student interest
	Content Understanding	Degree of teacher's ability to deepen students' understanding of key concepts, frameworks, and ideas
	Analysis and Inquiry	Degree of teacher's use of higher-level thinking strategies (e.g., support for problem solving, meta-cognition, novel application)
	Quality of Feedback	Degree of teacher and/or peer feedback used to expand learning and understanding through participation
	Instructional Dialogue	Degree of teacher's use of purposeful and cumulative questioning to guide and prompt depth of discussion/learning
	Student Engagement	Degree of students' active engagement in classroom activities

Table 3.2.

Brief information (location, sample size, grades), means, and standard deviations (in parentheses, if available) for studies using all dimensions of the published CLASS-S (with the exception of Student Engagement).

	Cannata, Smith, & Haynes, 2017	Smith et al., unpublished	Virtanen et al., 2018	Westergård, Ertesvåg, & Rafaelen, 2018	Yoder, 2013
Location of Study	USA	Canada	Finland	Norway	USA
Sample Size (Teachers)	N = 72	N = 38	N = 46	N = 52	N = 19
Grade Range	Grades 9-10	Grades 7-8	Grade 6	Secondary grades	Grades 4-8
Emotional Support	4.27	4.90 *(0.93)*	4.38	4.62 *(0.78)*	4.34 *(0.85)*
Positive Climate	4.82	5.38 *(1.06)*	4.93	4.99 *(0.89)*	4.58 *(0.97)*
Teacher Sensitivity	4.74	5.13 *(1.05)*	5.14	5.04 *(0.93)*	4.00 *(0.97)*
Regard for Adolescent Perspective	3.26	4.17 *(1.27)*	3.07	3.83 *(0.76)*	2.65 *(1.13)*
Classroom Organization	5.71	5.88 *(0.72)*	6.15	6.07 *(0.70)*	4.99 *(1.00)*
Behavioral Management	5.44	5.74 *(1.01)*	5.79	6.05 *(0.94)*	5.20 *(1.08)*
Productivity	5.34	5.27 *(1.04)*	5.89	4.52 *(0.89)*	5.41 *(0.98)*
Negative Climate (reversed)	6.36	6.62 *(0.63)*	6.78	6.65 *(0.45)*	6.12 *(0.66)*
Instructional Support	3.80	3.85 *(0.93)*	3.37	3.36 *(0.75)*	2.83 *(1.00)*
Instructional Learning Formats	4.55	4.55 *(1.02)*	4.80	4.60 *(0.94)*	4.35 *(1.11)*
Content Understanding	4.52	4.09 *(1.20)*	3.84	3.72 *(0.90)*	3.22 *(1.02)*
Analysis and Inquiry	2.70	3.29 *(1.22)*	2.46	2.38 *(0.79)*	2.64 *(1.05)*
Quality of Feedback	3.91	3.88 *(1.09)*	3.02	3.44 *(0.83)*	3.02 *(1.08)*
Instructional Dialogue	3.30	3.43 *(1.30)*	2.73	2.66 *(0.92)*	2.44 *(1.20)*
*Student Engagement**	4.67	5.01 *(1.19)*			

*Student Engagement scores were only available for two studies. Note: Negative Climate has been reverse scored. Domain scores based on the CLASS – Secondary Manual (Pianta, Hamre, & Mintz, 2012).

Table 3.3.

Descriptive statistics for all CLASS-S domains and dimensions.

Domain or Dimension	M	SD	Skew	SE skew	Kurt	SE kurt	Min	Max
Emotional Support[a]	4.90	0.93	-0.26	0.14	-0.25	0.29	2.00	7.00
Positive Climate	5.38	1.06	-0.52	0.14	-0.07	0.29	2.00	7.00
Teacher Sensitivity	5.13	1.05	-0.37	0.14	-0.39	0.29	2.00	7.00
Regard for Adolescent Perspective	4.17	1.27	-0.002	0.14	-0.63	0.29	1.00	7.00
Classroom Organization[a]	5.88	0.72	-1.02	0.14	0.81	0.29	3.00	7.00
Behavioral Management	5.74	1.01	-0.90	0.14	0.64	0.29	2.00	7.00
Productivity	5.27	1.04	-0.93	0.14	1.00	0.29	1.00	7.00
Negative Climate (reversed)	6.62	0.63	-1.67	0.14	2.65	0.29	4.00	7.00
Instructional Support[a]	3.85	0.93	0.14	0.14	-0.51	0.29	1.40	6.40
Instructional Learning Formats	4.55	1.02	-0.15	0.14	-0.64	0.29	2.00	7.00
Content Understanding	4.09	1.20	-0.01	0.14	-0.40	0.29	1.00	7.00
Analysis and Inquiry	3.29	1.22	0.16	0.14	-0.46	0.29	1.00	7.00
Quality of Feedback	3.88	1.09	0.38	0.14	-0.40	0.29	1.00	7.00
Instructional Dialogue	3.43	1.30	0.12	0.14	-0.45	0.29	1.00	7.00
Student Engagement	5.01	1.19	-0.45	0.14	-0.40	0.29	1.00	7.00

Note: M = mean, SD = standard deviation, Skew = skewness, SE skew = standard error for skewness, Kurt = kurtosis, SE kurt = standard error kurtosis, Min = minimum score, Max = maximum score.

[a]Domain scores based on the CLASS – Secondary Manual (Pianta, Hamre, & Mintz, 2012).

Table 3.4.

Correlations among CLASS-S domains and dimensions.

Scale	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.
Domains[a]														
1. Emotional Support	0.47**	0.59**	0.65**	0.72**	0.68**	0.49**	0.57**	0.47**	0.69**	0.65**	0.72**	0.79**	0.73**	0.57**
2. Classroom Organization		0.35**	0.48**	0.49**	0.21**	0.87**	0.83**	0.66**	0.30**	0.26**	0.28**	0.34**	0.23**	0.52**
3. Instructional Support			0.37**	0.49**	0.58**	0.21**	0.39**	0.22**	0.73**	0.77**	0.81**	0.85**	0.82**	0.37**
Dimensions														
4. Positive Climate				0.68**	0.47**	0.40**	0.29**	0.53**	0.31**	0.17**	0.25**	0.41**	0.34**	0.53**
5. Teacher Sensitivity					0.43**	0.40**	0.40**	0.38**	0.42**	0.28**	0.33**	0.54**	0.38**	0.60**
6. Regard for Adolescent Perspective						0.16**	0.16**	0.20**	0.48**	0.36**	0.52**	0.43**	0.50**	0.37**
7. Behavioral Management							0.54**	0.46**	0.17**	0.14*	0.18**	0.22**	0.14*	0.44**
8. Productivity								0.30**	0.34**	0.31**	0.29**	0.34**	0.27**	0.42**
9. Negative Climate (reversed)									0.19**	0.15*	0.20**	0.25**	0.13*	0.38**
10. Instructional Learning Formats										0.50**	0.49**	0.50**	0.48**	0.38**
11. Content Understanding											0.50**	0.61**	0.48**	0.22**
12. Analysis and Inquiry												0.62**	0.62**	0.29**
13. Quality of Feedback													0.67**	0.38**
14. Instructional Dialogue														0.24**
15. Student Engagement														

Note: Negative Climate has been reverse scored. *p <0.05 **p <0.01 (type two-tailed)

[a]Domain scores based on the CLASS – Secondary Manual (Pianta, Hamre, & Mintz, 2012).

Table 3.5.

Inter-rater reliability: Percent agreement (points within one [PWO]) and intra-class correlation coefficients (ICC).

Dimension	% Agreement PWO	ICC
Positive Climate	91.5	.67**
Teacher Sensitivity	86.4	.55**
Regard for Adolescent Perspective	91.5	.79**
Behavioral Management	81.4	.43*
Productivity	74.6	.32
Negative Climate	98.3	.40*
Instructional Learning Formats	88.1	.67**
Content Understanding	76.3	.64**
Analysis and Inquiry	69.5	.31
Quality of Feedback	74.6	.62**
Instructional Dialogue	64.4	.41*
Student Engagement	89.8	.67**
Overall	***82.2***	

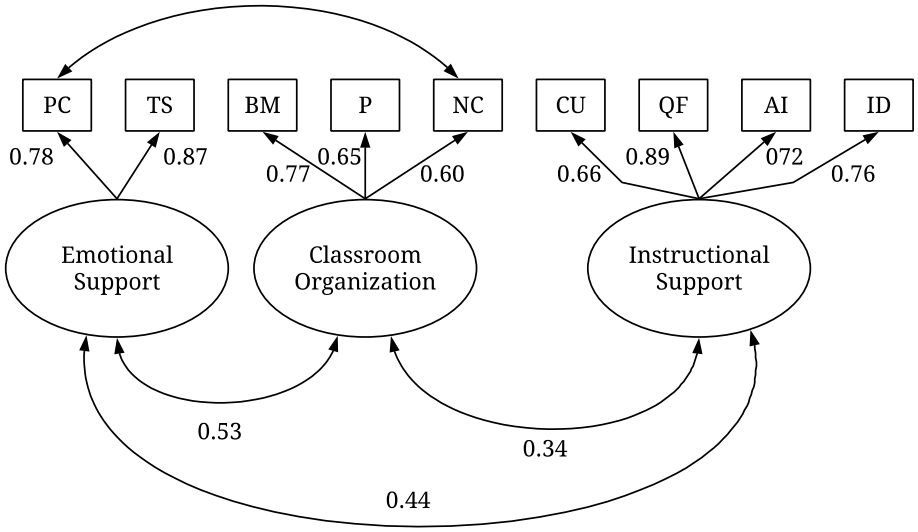
** $p < 0.05$, ** $p < 0.001$*

Table 3.6.

Model fit indices for various confirmatory factor models.

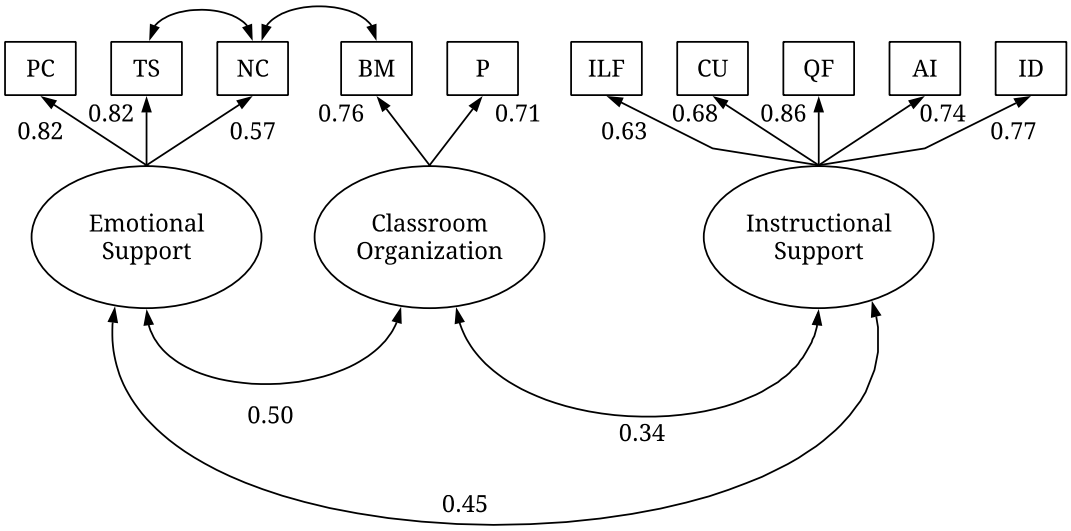
	X^2(df)	CFI	TLI	RMSEA	SRMR
CFA Models (at the rating level)					
Three-Factor Model (e.g., Hafen et al., 2015)	231.92(41)	0.86	0.81	0.13	0.08
One-Factor Model (e.g., McCaffrey et al., 2015)	457.72(44)	0.70	0.62	0.18	0.11
Modified Three-Factor Model[a] (e.g., Virtanen et al., 2018)	77.73(23)	0.95	0.92	0.09	0.05
Alternate Three-Factor Model[b]	101.56(30)	0.94	0.91	0.09	0.05
CFA Models (at the classroom level)					
Three-Factor Model (e.g., Hafen et al., 2015)	83.82(41)	0.84	0.78	0.17	0.10
One-Factor Model (e.g., McCaffrey et al., 2015)	134.62(44)	0.66	0.57	0.24	0.13
Modified Three-Factor Model[a] (e.g., Virtanen et al., 2018)	35.98(23)	0.94	0.91	0.12	0.08
Two-Factor Model[c]	70.59(33)	0.86	0.81	0.17	0.09

Note: X^2(df) = Chi square with degree of freedom; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Squared Error of Approximation; SRMR = Standardized Root Mean Square Residual (in the multilevel models the SRMR within); Modified Three Factor Model[a] = Regard for Adolescent Perspective and Instructional Learning Formats excluded, and Negative Climate and Positive Climate allowed to correlate; Alternate Three-Factor Model[b] = Negative Climate moved to the Emotional Support domain, Regard for Adolescent Perspective excluded, and residuals of Negative Climate and Teacher Sensitivity, and Negative Climate and Behavior Management allowed to correlate; Two-Factor Model[c] = Classroom Organization as one domain, and all other items loaded on a second factor (i.e., Teaching Capacity), with Positive Climate and Teacher Sensitivity allowed to correlate.



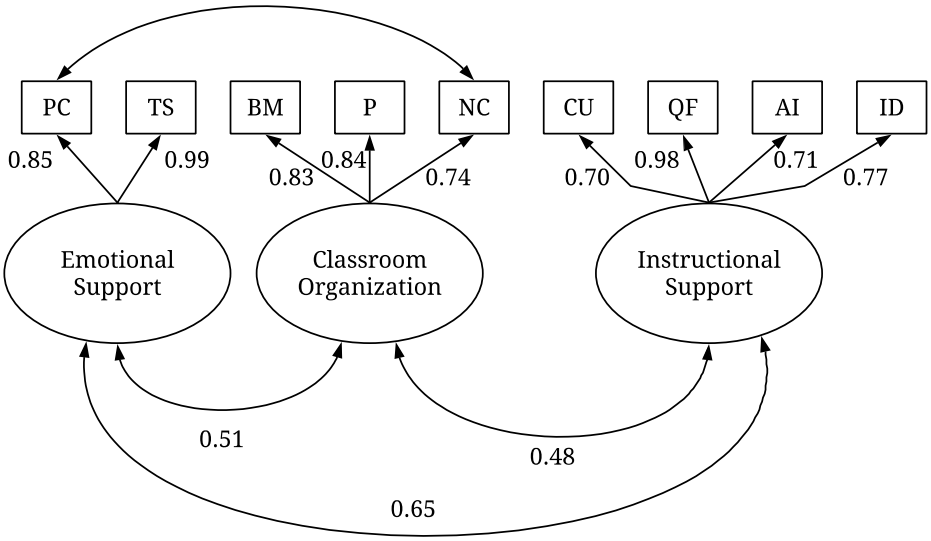
$X^2(23) = 77.73$, $p < 0.001$; CFI = 0.95, TLI = 0.92; RMSEA = 0.09; SRMR = 0.05

Figure 3.1. Modified three-factor model: Confirmatory factor analysis at the rating level. Modified three factor model (based on Virtanen et al., 2018) = Regard for Adolescent Perspective and Instructional Learning Formats excluded, and Negative Climate and Positive Climate allowed to correlate.



$X^2(30) = 101.56$, $p < 0.001$; CFI = 0.94, TLI = 0.91; RMSEA = 0.09; SRMR = 0.05

Figure 3.2. Alternative three-factor model: Confirmatory factor analysis at the rating level. Alternate three-factor model = Negative Climate moved to the Emotional Support domain, Regard for Adolescent Perspective excluded, and residuals of Negative Climate and Teacher Sensitivity, and Negative Climate and Behavior Management allowed to correlate.



$X^2(23) = 35.98$, p = 0.05; CFI = 0.94, TLI = 0.91; RMSEA = 0.12; SRMR = 0.08

Figure 3.3. Modified three-factor model: Confirmatory factor analysis at the classroom level. Modified three factor model (based on Virtanen et al., 2018) = Regard for Adolescent Perspective and Instructional Learning Formats excluded, and Negative Climate and Positive Climate allowed to correlate.

References

Alberta Education. (2010). Inspiring education: A dialogue with Albertans. Retrieved from https://open.alberta.ca/dataset/45370ce9-3a90-4ff2-8735-cdb760c720f0/resource/2ee2452c-81d3-414f-892f-060caf40e78e/download/4492270-2010-inspiring-education-dialogue-albertans-2010-04.pdf

Alberta Education. (2013). *Learning and technology policy framework.* Retrieved from https://education.alberta.ca/media/1046/learning-and-technology-policy-framework-web.pdf

Alberta Education. (2018). *Provincial achievement tests: PAT overview*. Retrieved from https://education.alberta.ca/provincial-achievement-tests/about-the-pats/

Allen, J. P., Hafen, C. A., Gregory, A. C., Mikami, A. Y., & Pianta, R. (2015). Enhancing secondary school instruction and student achievement: Replication and extension of the My Teaching Partner-Secondary intervention. *Journal of Research on Educational Effectiveness, 8*(4), 475-489. doi: 10.1080/19345747.2015.1017680

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333*(6045), 1034-1037. doi:10.1126/science.1207998

Allen, J., Gregory, A., Mikami, A., Lun, J. Hamre, B. & Pianta, R. (2013). Observations of effective teacher-student interactions in secondary school classrooms: Predicting student achievement

with the Classroom Assessment Scoring System-Secondary. *School Psychology Review, 42*(1), 76-97.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). AERA, APA, & NCME. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*, 62-87. doi:10.1080/10627197.2012.715014

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., McCaffrey, D. F., Gitomer, D. H., & Pianta, R. C. (2014). Improving observational score quality: Challenges in observer thinking. In T. J. Kane, K. A. Kerr & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project* (pp. 50-97). Retrieved from http://janbri.nl/wp-content/uploads/2015/01/Designing-Teacher-Evaluation-Systems_freePDF.pdf

Booker, L. N. (2014). *Examining the development of beginning middle school math teachers' practices and their relationship with the teachers' effectiveness* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Global. Vanderbilt University, Tennesee.

Cannata, M. A., Smith, T. M., & Taylor Haynes, K. (2017). Integrating academic press and support by increasing student ownership and responsibility. *AERA Open, 3*(3), 1-13. doi: 2332858417713181.

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement, 75*(2), 311-337. doi: 10.1177/0013164414539163

Caspersen, J., & Raaen, F. D. (2014). Novice teachers and how they cope. *Teachers and Teaching, 20*(2), 189-211. doi: 10.1080/13540602.2013.848570

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284. doi: 10.1037/1040-3590.6.4.284

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation, 10*(7), 1-9. Retrieved from http://www.pareonline.net/pdf/v10n7.pdf

Coughlan, S. (2017, August 2). How Canada became an education superpower. *BBC News.* Retrieved from http://www.bbc.com/news/business-40708421

Darling-Hammond, L., & Rothman, R. (2015). Developing effective teaching: Why search globally? In L. Darling-Hammond, & R. Rothman (Eds.), *Teaching in the Flat World: Learning from*

High-Performing Systems (pp. 1-5). New York: Teachers College Press.

Field, A. (2016). *Exploring data: The beast of bias.* Retrieved from https://www.discoveringstatistics.com/repository/exploringdata.pdf

Friedman, S. (2014). *Teacher emotional intelligence and the quality of interactions with students* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Global. Rutgers University, New Jersey.

Gamlem, S. M., & Munthe, E. (2014). Mapping the quality of feedback to support students' learning in lower secondary classrooms. *Cambridge Journal of Education, 44*(1), 75-92. doi: 10.1080/0305764X.2013.855171

Gill, B., Shoji, M., Coen, T., & Place, K. (2016). The content, predictive power, and potential bias in five widely used teacher observation instruments. *Mathematica Policy Research.* Retrieved from Regional Educational Laboratory website at http://ies.ed.gov/ncee/edlabs.

Gitomer, D., Bell, C., Qi, Y., McCaffrey, D., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record, 116*(060304), 1-32.

Government of Alberta. (2010). *Inspiring action on education.* Retrieved from https://www.oecd.org/site/eduilebanff/48763522.pdf

Gregory, A., Allen, J. P., Mikami, A. Y., Hafen, C. A., & Pianta, R. C. (2014). Effects of a

professional development program on behavioral engagement of students in middle and high school. *Psychology in the Schools, 51*(2), 143-163.

Hafen, C. A., Hamre, B. K., Allen, J. P., Bell, C. A., Gitomer, D. H., & Pianta, R. C. (2015). Teaching through interactions in secondary school classrooms: Revisiting the factor structure and practical application of the Classroom Assessment Scoring System–Secondary. *The Journal of Early Adolescence, 35*(5-6), 651-680. doi: 10.1177/0272431614537117

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials In Quantitative Methods For Psychology, 8*(1), 23. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3402032/

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development, 76*(5), 949-967. doi: 0009-3920/2005/7605-0001

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. doi: 10.1080/10705519909540118

Huang, F. L. (2017). *Conducting multilevel confirmatory factor analysis using R.* doi: 10.13140/RG.2.2.12391.34724

Joe, J. N. McClellan, C. A., & Holtzman, S. L. (2014). Scoring design decisions: Reliability and the

length and focus of classroom observations. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidelines from the Measures of Effective Teaching project* (pp. 415-443). San Francisco: Josey-Bass. Retrieved from http://k12education.gatesfoundation.org/wp-content/uploads/2015/11/Designing-Teacher-Evaluation-Systems_freePDF.pdf

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Research Paper. MET Project. Bill & Melinda Gates Foundation. Retrieved from http://files.eric.ed.gov/fulltext/ED540960.pdf

Lavigne, A. L., & Good, T. L. (2013). *Teacher and student evaluation: Moving beyond the failure of school reform.* New York: Routledge.

Malmberg, L.-E., & Hagger, H. (2009). Changes in student teachers' agency beliefs during a teacher education year, and relationships with observed classroom quality, and day-to-day experiences. *British Journal of Educational Psychology, 79*, 677-694. doi: 10.1348/000709909X454814

Malmberg, L.-E., Hagger, H., Burn, K., Mutton, T., & Colls, H. (2010). Observed classroom quality during teacher education and two years of professional practice. *Journal of Educational Psychology, 102*(4), 916-932. doi: 10.1037/a0020920

Mashburn, A. J., Meyer, J. P., Allen, J. P., & Pianta, R. C. (2014). The effect of observation length

and presentation order on the reliability and validity of an observational measure of teaching quality. *Educational and Psychological Measurement, 74*(3), 400-422. doi: 10.1177/0013164413515882

McCaffrey, D. F., Yuan, K., Savitsky, T. D., Lockwood, J. R., & Edelen, M. O. (2015). Uncovering multivariate structure in classroom observations in the presence of rater errors. *Educational Measurement: Issues and Practice, 34*(2), 34-46. doi: 10.1111/emip.12061

The Organization for Economic Co-Operation and Development. (2015). *Programme for International Student Assessment (PISA).* Retrieved from http://www.oecd.org/pisa/

Patton, M. Q. (2011). *Developmental evaluation: Applying complexity concepts to enhance innovation and use.* New York: Guildford Press.

Pianta, R. C. (2006). Classroom management and relationships between children and teachers: implications for research and practice. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 685-709). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109-119.

Pianta, R. C., Hamre, B. K., & Mintz, S. (2012). *Classroom Assessment Scoring System: Secondary*

manual. Charlottesville, VA: Teachstone.

Preston, C., Goldring, E., Guthrie, J. E., Ramsey, R., & Huff, J. (2017). Conceptualizing essential components of effective high schools. *Leadership and Policy in Schools, 16*(4), 525-562. doi: 10.1080/15700763.2016.1205198

Puentadura, R. (2010). *Substitution, Adaptation, Modification, and Redefinition Model.* Retrieved from: http://www.hippasus.com/rrpweblog/archives/2012/08/23/SAMR_BackgroundExemplars.pdf

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

RStudio Team (2018). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. Retrieved from http://www.rstudio.com/.

Rose, D.H. & Meyer, A. (2006). *A Practical Reader in Universal Design for Learning.* Cambridge, MA: Harvard Education Press.

Rosseel, Y. (2017). lavaan: An R package for structural equation modeling. Journal of Statistical Software, 48(2), 1-36. Retrieved from http://www.jstatsoft.org/v48/i02/

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68. doi: 10.1037110003-066X.55.1.68

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods Of Psychological Research Online, 8*(2), 23-74. Retrieved from http://citeseerx.ist.psu.edu.login.ezproxy.library.ualberta.ca/viewdoc/download?doi=10.1.1.509.4258&rep=rep1&type=pdf

Shyamalan, M. N. (2013). *I Got Schooled: The Unlikely Story of How a Moonlighting Movie Maker Learned the Five Keys to Closing America's Education Gap*. New York: Simon and Schuster.

Smith, V., Hayes, S., Labonte, C., & Vargas, F. (2016). *Flexible pathways to success: Technology to design for diversity.* Retrieved from the Government of Alberta website: https://open.alberta.ca/dataset/029047d9-3a06-4b18-9c94-40d0c3937639/resource/e67821ad-ce9a-4f73-9248-b9710ebd6cf2/download/uofa-flexible-pathways-finalreport-july26-2016-online.pdf

Vaske, J. J., Beaman, J., & Sponarski, C. C. (2017). Rethinking internal consistency in Cronbach's alpha. *Leisure Sciences, 39*(2), 163-173. doi: 10.1080/01490400.2015.1127189

Virtanen, T. E., Lerkkanen, M.-K., Poikkeus, A.-M., & Kuorelahti, M. (2015). The relationship between classroom quality and students' engagement in secondary school. *Educational Psychology: An International Journal of Experimental Educational Psychology, 35*(8), 963-

983. doi: 10.1080/01443410.2013.822961

Virtanen, T. E., Pakarinen, E., Lerkkanen, M. K., Poikkeus, A. M., Siekkinen, M., & Nurmi, J. E. (2018). A validation study of Classroom Assessment Scoring System–Secondary in the Finnish school context. *The Journal of Early Adolescence, 38*(6), 849-880. doi: DOI: 10.1177/0272431617699944

World Education Forum. (2015). Incheon Declaration Education 2030: Towards inclusive and equitable quality education and lifelong learning for all. Retrieved from http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/ED/ED/pdf/FinalVersion-IncheonDeclaration.pdf

Wragg, T., & Wragg, E. C. (2012). *An Introduction to Classroom Observation.* New York, NY: Routledge.

Yoder, N. P. (2013). *Linking social-emotional learning to teaching practices in an urban context: A mixed-methods study* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Global. (UMI 3610176).

Yuan, K., McCaffrey, D. F., & Savitsky, T. D. (2013). *Analyzing the factorial structure of the Classroom Assessment Scoring System-Secondary using a Bayesian hierarchical multivariate ordinal model.* Paper presented at the Society for Research on Educational Effectiveness, Washington, DC. Retrieved from http://files.eric.ed.gov/fulltext/ED563074.pdf

Chapter Four. General Discussion and Conclusion

General Discussion

In order to support educational transformation, we need to deepen our understanding of the complexities of teaching and what constitutes effective classroom practices. Despite a long history of research examining effective teaching, debate remains regarding how to best conceptualize this construct (Stronge, Ward, & Grant, 2011). Traditionally, researchers have focused on student achievement (e.g., results of standardized tests), instructional practices (e.g., questioning), and teacher characteristics (e.g., qualifications) as indicators of effective classroom practices (Stronge et al., 2011). However, advances in methodologies, technology, and contributions from large-scale research projects have provided opportunities to enrich our understanding of complex classroom environments (Kane & Staiger, 2012; Klette & Blikstad-Balas, 2018; McCombs, 2012; Stronge et al., 2011). One method that has been increasingly supported by researchers is the use of standardized observational protocols in the classroom. In particular, evidence supporting the use of the CLASS-S as a reliable and valid observational tool for use in secondary school classrooms has emerged in literature from the United States (e.g., Hafen, Hamre, Allen, Bell, Gitomer, & Pianta, 2015), United Kingdom (Malmberg, Hagger, Burn, Mutton, & Colls, 2010), Finland (e.g., Virtanen, Pakarinen, Lerkkanen, Poikkeus, Siekkinen, & Nurmi, 2018), and Norway (Westergård, Ertesvåg, & Rafaelsen, 2018).

Furthermore, researchers have described that higher quality classroom practices, as measured by the CLASS-S, are associated with various positive student outcomes (e.g., Allen, Pianta, Gregory, Mikami, & Lun, 2011; Culp, Martin, Clements, & Presser, 2015; Kane & Staiger, 2012; Virtanen, Lerkkanen, Poikkeus, & Kuorelahti, 2015). Therefore, the CLASS-S may be a valuable tool to deepen our understanding of the complexities of teaching and effective classroom practices, in order to better support educational transformation.

Educational Transformation and Educational Psychology

As previously noted, an international focus has emerged emphasizing the need for educational transformation to meet the rapidly changing needs of education for the 21st century (e.g., Government of Alberta, 2010; McCombs, 2012; O'Sullivan & Dallas, 2017; Tondeur, Aesaert, Pynoo, Braak, Fraeyman, & Erstad, 2017). In the 21st century, "content is so abundant as to make it a poor foundation for basing an educational system" (McCombs, 2012, p. 498). Therefore, educational transformation requires education (including administrators, teachers, researchers, and policy makers) to shift from traditional indicators of effective classroom practices, such as measures of content learning (i.e., student achievement), to expand our understanding of more holistic classroom practices (McCombs, 2012). For example, researchers have proposed the purpose of education in the 21st century is to teach students to use communication, collaboration, creativity, problem-solving, and critical thinking skills

(McCombs, 2012; Van Laar, van Deursen, van Dijk, & de Haan, 2017). Therefore, the applied science of educational psychology may be the field best positioned to drive educational transformation through research and methods that help bridge what we know to support shifts in classroom practices (McCombs, 2012). For example, educational psychology recognizes the complexities of the interactions between systems and individuals, and researchers have identified a variety of environmental supports vital to secondary classrooms (McCombs, 2012; Shernoff, Ruzek, & Sinha, 2017). In brief, researchers have reported classroom environments should be responsive, and provide opportunities for autonomy, self-expression, competence, and supportive relationships between teachers and students, as well as peers (Shernoff et al., 2017). Importantly, these key features of classroom environments are captured by the CLASS-S.

Therefore, the CLASS-S may be a valuable tool for researchers in educational psychology, as well as school psychologists. School psychologists are often tasked with the role of bridging research to practice in educational systems. This role is consistent with practice standards outlined by the National Association of School Psychologists (NASP, 2010) indicating school psychologists' role includes supporting teachers to increase their skills through evidence-based practices. Drawing on evidence-based assessment, such as findings from the CLASS-S observational protocol, and the growing body of available research evidence, school psychologists are well positioned to support teachers' classroom practices to increase emotional

support, instructional practices, classroom organization, and student engagement (Shernoff et al., 2017). Furthermore, school psychologists may provide consultation to teachers, administrators, and policymakers to inform educational transformation by advising of research evidence, and new methods, such as the CLASS-S in order to clearly draw connections between what teachers do in the classroom and student outcomes (NASP, 2010; Shernoff et al., 2017). In addition, learning environments not only provide a deeper understanding of the complexities of the classroom, they are easier and quicker to assess than content learning (Shernoff et al., 2017).

What teachers do in the classroom and how they can be best supported through interventions has often been overlooked in research (Putnam & Borko, 2000). Those who provide professional development to teachers, such as school psychologists, have long struggled with how to create learning opportunities that are "powerful enough to transform teachers' classroom practice" (Putnam & Borko, 2000, p. 6). In order to meaningfully impact quality teaching practices in the classroom, activities must be relevant to the individual and their classrooms. Therefore, one of the most valuable applications of standardized classroom observational systems is to guide and support professional development (Danielson, 2011; Pianta & Hamre, 2009). "The true promise of classroom observations is the potential to identify strengths and address specific weaknesses in teachers' practice" (Kane & Staiger, 2012, p.14).

Summary of Contributions from the Present Research

Despite a myriad of published research using various versions of the CLASS, the present work represents two unique contributions to the literature based on the CLASS-S: (1) Chapter Two represents the first known scoping review to identify, map, summarize, and provide recommendations specific to the CLASS-S; and (2) Chapter Three represents the first known preliminary reliability and validity study of the CLASS-S in Alberta, Canada. Notably, only one other review article has been identified in the literature, and it focuses on the CLASS in Early Childcare and Education programs for children aged 30 to 72 months (Perlman, Falenchuk, Fletcher, McMullen, Beyene, & Shah, 2016). Furthermore, no known studies have reported on the preliminary reliability and validity of the CLASS-S in any Canadian province. The only known studies conducted outside of the United States to report on the reliability and validity of the CLASS-S have been from the United Kingdom (Malmberg et al., 2010), Finland (Virtanen et al., 2018) and Norway (Westergård et al., 2018).

The present research provided an opportunity to summarize what is known with respect to the CLASS-S, and to identify what research is still needed. The following synopsis of Chapters Two and Three will present themes based on integrated findings from across chapters, while summarizing what was learned, and what remains to be learned about the CLASS-S. Themes addressed in the present research included the psychometric properties of the CLASS-S, associations between quality classroom practices (as defined by the Teaching Through

Interactions Framework and the CLASS-S) and student outcomes, and the emerging use of the CLASS-S in international research.

Psychometric properties: What was learned. One consistent theme discussed across Chapters Two and Three was findings specific to the psychometric properties of the CLASS-S, particularly those related to reliability and validity. Notably, the present research identified that the majority of authors using the CLASS-S reported following standard training certification procedures, meaning that certified coders received training in the use of the CLASS-S from the developers, a process that supports reliability. This highlights a strength of the CLASS-S as compared to other standardized observational protocols (e.g., the Framework for Teaching; Danielson, 2011), as its reliability is supported by consistent training and certification (and in turn re-certification) procedures for coding. Although authors often reported inter-rater reliability, the present study highlighted methodological concerns about how these statistics were calculated. In particular, the practice of relying on agreement within one-point as an indicator of inter-rater reliability is particularly flawed (Hallgren, 2012). Furthermore, Burchinal (2018) suggested a one-point discrepancy between coders typically reflects a difference of one standard deviation, thus attributing a great deal of measurement error to the differences between coders. In fact, some researchers have identified rater drift and discrepancy as a significant source of error (Casabianca, Lockwood, & McCaffrey, 2015).

With regards to the validity of the CLASS-S, according to the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014), an accumulation of support is required for every interpretation. Some relevant categories of evidence to support the validity of the CLASS-S include validity of structure, content, relationships with conceptually related constructs, and relationships with criteria. Importantly, at this time the majority of studies have not used the published version of the CLASS-S. As such, caution is warranted when making broad conclusions about the psychometrics of the tool, as there has not been enough research using the finalized tool, across the wide range of ways it has been used (e.g., as an indicator of quality classroom teaching, as a means for accountability to assess and provide feedback to teachers, as a way to compare diverse settings, or as a model to guide interventions).

However, one category of evidence received substantial attention in the present research, and that was evidence specific to the validity of structure. Although it should be noted various publications explored the factor structure of the CLASS-S (e.g., Allen, Gregory, Mikami, Lun, Hamre, & Pianta, 2013; Hafen et al., 2015; Malmberg, et al., 2010; McCaffrey et al., 2015; Virtanen et al., 2018; Westergård et al., 2018), many studies were not based on the finalized version of the CLASS-S and so may not have included the same dimensions. Nevertheless, as

reported in the present research, the general three-factor model of the CLASS-S appears to hold up across many studies and educational contexts, with some variations reported to strengthen models of fit. Although the validity of the factor structure has been reported as falling in the acceptable range in published literature, many of these models fall in the lower range of what is statistically considered to be a good fitting model. For example, some models (including the model identified in Chapter Three) proposed reducing one of the factors (i.e., domains) to include only two dimensions. Commonly, most researchers recommend a minimum of three dimensions per factor in order to ensure a stronger model (Costello & Osborne, 2005).

Psychometric properties: What needs to be learned. Further exploration is needed to solidify our knowledge of the psychometric properties of the CLASS-S. In particular, more studies exploring sources of error, and procedures to reduce such errors (e.g., double-coding, number of segments, timeframes/schedules for observations), based on the published version of the CLASS-S would be beneficial to help understand the strengths and limitations of the tool. For example, knowing how classroom practices based on the CLASS-S change over the course of a year, and across multiple years, may help us better understand patterns and optimal times for observation, and intervention. In addition, more complex inter-rater reliability statistics should be included in future studies, such as intra-class correlation coefficients.

Moreover, further validity studies exploring the wide range of evidence as outlined by the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) for each potential use of the CLASS-S is needed. Specifically, more should be learned about the relationship between the CLASS-S and underlying theories of adolescent development, similar quality classroom teaching constructs, and with the criteria it is intended to measure (e.g., effective teaching, student achievement). Despite the range of articles addressing the factor structure, large scaled studies using the final version of the CLASS-S across a variety of contexts (e.g., types of classrooms, countries) are needed. Due to the inherent nature of the classroom observation data (i.e., nested structure at the individual rating level and the aggregated classroom level), further explorations of the factor structure should include bi-factor and multilevel factor analyses in an effort to account for the complexities of measuring classroom observations. Studies drawing from larger sample sizes would allow researchers to fully explore the factor structure using more advanced statistical methods (Huang, 2017).

Associations between the CLASS-S and student outcomes: What was learned. In an extension of the discussion of the validity of the CLASS-S, one theme from this research related to evidence supporting the use of the CLASS-S as an indicator of effective classroom practices, and specifically, its relationship to the criteria of positive student outcomes. In other words, if the CLASS-S captures effective classroom practices, consistent evidence linking the CLASS-S

to positive student outcomes (e.g., student achievement, student engagement, and student perceptions) should be available in the literature. What was learned in the present research is that evidence is emerging to support the association between the CLASS-S and some student outcomes. In brief, researchers have identified higher overall CLASS-S, or Emotional Support domain scores, in high track (or honors track) classrooms as compared to regular track classrooms (Donaldson, LeChasseur, & Mayer, 2017; Rutledge, Cohen-Vogel, & Osborne-Lampkin, 2012; Smith, Preston, Haynes, & Neergaard Booker, 2015). These findings suggest an association between classroom quality, as measured by the CLASS-S, and improved student performance.

However, in studies exploring the relationship between the CLASS-S and student achievement, researchers have reported mixed findings. As previously noted, methods used to measure student achievement (e.g., state mandated subject or achievement tests, GPAs, value-added scores) in the literature, have been criticized. In Chapter Two, only one article identified in the scoping review was designed specifically to assess the relationship between the CLASS-S and student achievement. The researchers reported that end of year student achievement (based on a state mandated subject test) was predicted by all three CLASS-S domains, even after controlling for previous achievement scores (Allen et al., 2013). They identified the Emotional Support domain had the strongest predictive value and concluded that greater Emotional Support

led to greater year-end gains in student achievement, despite prior achievement test scores. Similar findings were described by researchers exploring the impact of interventions (i.e., My Teaching Partner – Secondary [Allen et al., 2011; Allen, Hafen, Gregory, Mikami, & Pianta, 2015] and the Exploring Photosynthesis program [Culp, Martin, Clements, & Presser, 2015]), where higher scores on the CLASS-S were associated with greater gains in student achievement (as measured by state mandated achievement tests, or end of unit test scores). In contrast, a researcher used self-reported grade point averages (GPAs) as an indicator of student achievement, and did not identify any associations between the CLASS-S and student achievement (Yoder, 2013).

In other studies, researchers sought to measure student achievement using more complex scores, known as value-added models. Value-added models were used to estimate a teacher's contribution to student achievement by controlling other factors (e.g., prior achievement, class size, years of teaching experience). Importantly, value-added models vary based on the factors researchers identify as part of the model, and as such caution is warranted when relying on these models to make conclusions (e.g., AERA, 2015). In brief, researchers have found associations between CLASS scores and moderate gains in student achievement in a large-scale project based on individual value-added scores (Kane & Staiger, 2012; Ruzek, Hafen, Hamre, & Pianta, 2014). In contrast, when value-added scores were used to differentiate between high value-added and

low value-added schools, some studies reported no differences on the CLASS-S (Grossman, Loeb, Cohen, & Wyckoff, 2013; Rutledge et al., 2012, Smith et al., 2015), while others found higher Emotional Support (not including the Positive Climate dimension) in HVA schools (Smith, Cannata, & Haynes, 2016). These findings reflect the fact the CLASS-S likely captures more within school (i.e., at the classroom level), than between school variability.

Student engagement was another student outcome addressed in the present research. Importantly, the Student Engagement dimension of the CLASS-S is considered an outcome of the three broad domains of the CLASS-S, and in turn may be considered a student outcome. Furthermore, additional research described herein summarized emerging support for the association between the CLASS-S and other measures of student engagement. Measurement of student engagement included student-ratings, teacher-ratings, and observation scores from the CLASS-S. In a study comparing all three methods of measurement, student engagement was associated with higher CLASS-S scores across all domains regardless of method (Virtanen et al., 2015). Two studies using only student-ratings of student engagement found associations between this construct and the CLASS-S (Hafen et al., 2012; Ruzek et al., 2016), while one study using student-ratings did not identify any associations between student engagement and the CLASS-S (Yoder, 2013). Additionally, authors of one study using only the CLASS-S observational measure of student engagement concluded the association between student

engagement and the CLASS-S domains varied between the individual segment and aggregated classroom level (Malmberg et al., 2010). In particular, higher student engagement predicted higher CLASS-S domain scores at the segment level; however, when scores were aggregated across segments at the classroom level, only associations with the Classroom Organization domain remained. Findings summarized identified the Student Engagement dimension of the CLASS-S protocol provided a strong indicator of student engagement, beyond more traditional measures based on student- or teacher-ratings (e.g., Malmberg et al., 2010; Virtanen et al., 2015).

The last student outcome addressed in the present research was the association between the CLASS-S and student's perspectives of various classroom contextual factors (i.e., classroom structure, classroom climate). There was minimal evidence reported in this category, with some researchers' making links between various CLASS-S dimensions and students' perception of classroom structure (Spearman & Watt, 2013); while others did not find any links between the CLASS-S and students' reports of classroom climate (Yoder, 2013). Therefore, evidence related to student perspectives of classroom context reported in the present research was limited.

Associations between the CLASS-S and student outcomes: What needs to be learned. Despite documenting some emerging evidence for associations between the CLASS-S and student outcomes, specifically student achievement and student engagement, more research is needed to support the CLASS-S as an observational measure of effective classroom practices.

In the present research, links between classroom quality, as measured by the CLASS-S, and student ability were noted; however, questions about this relationship remain unanswered. For example, are teachers able to use more effective classroom practices when students already have pre-existing academic abilities (i.e., in high track or honours classrooms), or are other student-driven factors impacting the scores on the CLASS-S by providing more opportunities for raters to see and score more teacher-student interactions? Also, given the shifting needs of education for the 21st century to move away from an emphasis on learning content, it should be noted that student achievement was the most common student outcome measured in the literature and was notably frequently assessed by tests of content (either state standardized tests of achievement, or unit tests). Therefore, studies linking the CLASS-S to student outcomes that measure 21st century skills such as the ability to evaluate, analyze, and apply knowledge to solve novel problems (O'Sullivan & Dallas, 2017) is warranted.

Furthermore, caution is required when relying on value-added models to make conclusions about associations between the CLASS-S and effective classroom practices. Based on findings reported in this research, when value-added models have differentiated between schools, few to no differences were identified based on the CLASS-S. One finding noted in the present research was that the CLASS-S was associated with more variability at the classroom level than at the school level, likely representing more variability within schools than between

schools (Rutledge et al., 2012). Therefore, as recommended by the AERA (2015), more research exploring alternative methods and models for identifying effective classrooms is required, and in turn further research exploring the association between the CLASS-S and these other methods are warranted.

As noted, the Student Engagement dimension of the CLASS-S may be a valuable observational measure of student engagement in junior/middle or high school classrooms. Although support reported in the present research was limited (i.e., based on few studies), further research would be beneficial. In addition, more exploration linking the Student Engagement dimension to underlying theories of the CLASS-S, for example peer-relatedness, may be helpful. In Chapter Two findings from one study reported increased student engagement was associated with a larger class size (Malmberg et al., 2010), and therefore investigating if the CLASS-S indicators of student engagement reflect adolescent theories of development may address discrepancies from traditional reports linking student engagement with smaller class sizes.

Last, links between the CLASS-S and student perspectives of classroom context reported in this document were limited and inconclusive. More research triangulating students' perspectives of classroom context including, but not limited to, measures of classroom structure and classroom climate, are justified. Expanding research to investigate the evidence for the validity of the CLASS-S as linked to students' perspectives of effective classroom practices

would be helpful. Research drawing on mixed methods approaches may provide more information supporting the CLASS-S from the perspective of students.

The CLASS-S in international research: What was learned. A final theme addressed across Chapter Two and Three in the present research was the use of the CLASS-S in international contexts. Although the CLASS-S was first developed in the United States, other countries have adopted its use. In particular, articles originating from Australia, Finland, Norway, and the United Kingdom were presented in this document. However, few of these articles used the published version of the CLASS-S. One study conducted in the United Kingdom explored how classroom practices changed during teacher training and the first two years of their career (Malmberg et al., 2010). This study used an earlier iteration of the CLASS-S and researchers conducted a factor analysis that reduced the number of dimensions, and subsequently suggested a single-factor was the best indicator of classroom quality (Malmberg et al., 2010). Although study findings were detailed elsewhere in this document, one salient finding was that Classroom Organization domain scores increased linearly with teaching experience, and similar findings have been reported in the United States (Booker, 2014; Neergaard & Smith, 2012). Furthermore, Classroom Organization scores in general appear to be consistently scored higher than the other domains, while Instructional Support scores appear to be consistently

lower, regardless of the location where the study occurred. This suggests some CLASS-S scoring patterns may be stable across international contexts.

In other international studies, two recent articles explored the factor structure in Finland (Virtanen et al., 2018), and Norway (Westergård et al., 2018), and used the final published version of the CLASS-S. Additionally, the present study contributed the first international exploration of the factor structure of the CLASS-S in a sample of classroom observations from Alberta, Canada. In brief, results from these studies supported the use of the CLASS-S in Finland, Norway, and Alberta, Canada (see Chapter Three for details). Similar findings were reported across the Finnish, Norwegian, and Canadian studies, supporting the three-factor structure of the CLASS-S (with some minor variations), indicating it may be a valuable tool in each of these educational contexts. Notably, Finland, Norway, and Canada, are all leaders in education according to international rankings (The Organization for Economic Co-Operation and Development, 2015).

The CLASS-S in international research: What needs to be learned. Although some research has originated from Australia, Finland, Norway, the United Kingdom, and now Canada, more international research is required. Specifically, more evidence is first required to support the validity of the CLASS-S in each of these contexts. As previously summarized, the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) require evidence of

the validity of structure, content, relationships with conceptually related constructs, and relationships with criteria, and international studies should also seek to address these categories of evidence. For example, a limitation of the current research presented in Chapter Three, a preliminary exploration of reliability and structural validity of the CLASS-S in Alberta, Canada, is that additional categories of evidence supporting validity were not provided. For example, further research in Alberta, Canada, linking the CLASS-S to other related constructs, and criteria (e.g., peer relatedness, student outcomes, 21st century skills, theories of adolescent development, measures of effective classroom practices) would support its use in the Canadian context. Of the international studies of CLASS-S factor structure, one (Virtanen et al., 2018) also sought to link the CLASS-S to other constructs, and reported correlations between the Classroom Organization domain, and measures of classroom management efficacy beliefs, teaching-related stress, and teacher exhaustion.

Once adequate evidence is collected for the use of the CLASS-S in international contexts, it may also provide a means to make comparisons across these diverse educational settings. In particular, further research might determine universal practices of effective classroom teaching. Furthermore, the CLASS-S may be useful in other international contexts as a means to capture effective teaching practices and identify areas of practice in need of intervention. For example, in countries where educational systems are struggling to support students, the CLASS-S may be

a useful tool to provide feedback, guide teacher professional development, support universal practices of effective teaching, and drive international educational transformation.

Conclusion

Together, Chapter Two and Three highlighted that, despite the use of the CLASS-S in published research since 2009, the evidence base supporting its reliability, validity, and utility is still emerging. Importantly, the CLASS-S has the potential to contribute to our understanding of effective classroom practices from both the research and practice perspectives. The CLASS-S may serve as a means to investigate the most impactful classroom practices, and link them to student outcomes, so that teachers can be supported to meet the new demands required for 21st century educational transformation. Furthermore, the CLASS-S may be a useful tool to make international comparisons, and to identify classroom practices that transcend content delivery, to support other factors shown to impact student outcomes beyond academic ability. Therefore, the present dissertation calls for increased use of the CLASS-S in international research and practice, in order to deepen our knowledge of effective classroom practices, to clarify their associations with other constructs and student outcomes, and to further support educational transformation.

References

Alberta Education. (2010). *Inspiring education: A dialogue with Albertans*. Retrieved from https://open.alberta.ca/dataset/45370ce9-3a90-4ff2-8735-cdb760c720f0/resource/2ee2452c-81d3-414f-892f-060caf40e78e/download/4492270-2010-inspiring-education-dialogue-albertans-2010-04.pdf

Alberta Education. (2013). *Learning and technology policy framework.* Retrieved from https://education.alberta.ca/media/1046/learning-and-technology-policy-framework-web.pdf

Alberta Education. (2018). *Provincial achievement tests: PAT overview*. Retrieved from https://education.alberta.ca/provincial-achievement-tests/about-the-pats/

Allen, J. P., Hafen, C. A., Gregory, A. C., Mikami, A. Y., & Pianta, R. (2015). Enhancing secondary school instruction and student achievement: Replication and extension of the My Teaching Partner-Secondary intervention. *Journal of Research on Educational Effectiveness, 8*(4), 475-489. doi: 10.1080/19345747.2015.1017680

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333*(6045), 1034-1037. doi:10.1126/science.1207998

Allen, J., Gregory, A., Mikami, A., Lun, J. Hamre, B. & Pianta, R. (2013). Observations of

effective teacher-student interactions in secondary school classrooms: Predicting student achievement with the Classroom Assessment Scoring System-Secondary. *School Psychology Review, 42*(1), 76-97.

Amato, W. W. (2012). *English language learners' experiences in mixed-language, mixed-culture classrooms* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Global. (UMI Number: 3525023)

American Educational Research Association. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. Educational Researcher, 44(8), 448-452. doi: 10.3102/0013189X15618385

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). AERA, APA, & NCME. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Archer, J., Kerr, K. A., & Pianta, R. C. (2014). Why measure effective teaching? In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), Designing teacher evaluation systems: New guidelines from the Measures of Effective Teaching project (pp. 415-443). San Francisco: Josey-Bass. Retrieved from http://k12education.gatesfoundation.org/wp-content/uploads/2015/11/Designing-Teacher-Evaluation-Systems_freePDF.pdf

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology, 8*(1), 19-32. doi: 10.1080/1364557032000119616

Athanases, S. Z., Achinstein, B., Curry, M. W., & Ogawa, R. T. (2016). The promise and limitations of a college going culture: Toward cultures of engaged learning for low-SES Latina/o Youth. *Teachers College Record, 118*, 070306.

Athanases, S. Z., & de Oliveira, L. C. (2014). Scaffolding versus routine support for Latina/o youth in an urban school: Tensions in building toward disciplinary literacy. *Journal of Literacy Research, 46*(2), 263-299. doi: 10.1177/1086296X14535328

Battistich, V., Schaps, E., & Wilson, N. (2004). Effects of an elementary school intervention on students' "connectedness" to school and social adjustment during middle school. *Journal of Primary Prevention, 24*, 243–262. doi:10.1023/B:JOPP.0000018048.38517.cd

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*, 62-87. doi:10.1080/10627197.2012.715014

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., McCaffrey, D. F., Gitomer, D. H., & Pianta, R. C. (2014). Improving observational score quality: Challenges in observer thinking. In T. J. Kane, K. A. Kerr & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New*

guidance from the Measures of Effective Teaching project (pp. 50-97). Retrieved from http://janbri.nl/wp-content/uploads/2015/01/Designing-Teacher-Evaluation-Systems_freePDF.pdf

Berry, B. (2011). Teacherpreneurs: A more powerful vision for the teaching profession. *Phi Delta Kappan, 92*(6), 28-33. Retrieved from https://wyoming2011uwyo.pbworks.com/f/Teacher%20Entrepreneurs%20Barnett.pdf

Booker, L. N. (2014). *Examining the development of beginning middle school math teachers' practices and their relationship with the teachers' effectiveness* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Global. Vanderbilt University, Tennesee.

Brandt, C., Meyers, C., & Molefe, A. (2013). *The impact of eMINTS professional development on teacher instruction and student achievement: Year 1 Report.* Retrieved from eMINTS website http://emints.org/wp-content/uploads/2014/09/The-Impact-of-eMINTS-Professional-Development-on-Teacher-Instruction-and-Student-Achievement-%E2%80%93-Year-1-Report-January-2013.pdf

Cannata, M. A., Smith, T. M., & Taylor Haynes, K. (2017). Integrating academic press and support by increasing student ownership and responsibility. *AERA Open, 3*(3), 1-13. doi: 2332858417713181.

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement, 75*(2), 311-337. doi: 10.1177/0013164414539163

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement, 73*(5), 757-783. doi:10.1177/0013164413486987

Caspersen, J., & Raaen, F. D. (2014). Novice teachers and how they cope. *Teachers and Teaching, 20*(2), 189-211. doi: 10.1080/13540602.2013.848570

Cherng, H. Y. S., & Halpin, P. F. (2016). The importance of minority teachers: Student perceptions of minority versus White teachers. *Educational Researcher, 45*(7), 407-420. doi: 10.3102/0013189X16671718

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284. doi: 10.1037/1040-3590.6.4.284

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources, 41*(4), 778-820. doi: 10.3368/jhr.XLI.4.778

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation, 10*(7), 1-9. Retrieved from http://www.pareonline.net/pdf/v10n7.pdf

Coughlan, S. (2017, August 2). How Canada became an education superpower. *BBC News.* Retrieved from http://www.bbc.com/news/business-40708421

Culp, K. M., Martin, W., Clements, M., & Presser, A. L. (2015). Testing the impact of a pre-instructional digital game on middle-grade students' understanding of photosynthesis. *Technology, Knowledge and Learning, 20*(1), 5-26. doi: 10.1007/s10758-014-9233-5

Danielson, C. (2011). *Enhancing Professional Practice: A Framework for Teaching.* Alexandria, VA: ASCD.

Donaldson, M. L., LeChasseur, K., & Mayer, A. (2017). Tracking instructional quality across secondary mathematics and English Language Arts classes. *Journal of Educational Change, 18*(2), 183-207. doi: 10.1007/s10833-015-9269-x

Darling-Hammond, L., & Rothman, R. (2015). Developing effective teaching: Why search globally? In L. Darling-Hammond, & R. Rothman (Eds.), *Teaching in the Flat World: Learning from High-Performing Systems* (pp. 1-5). New York: Teachers College Press.

Downer, J. T., Jamil, F., Maier, M. F., & Pianta, R. C. (2012). Implications of information

processing theory for professional development of early educators. In C. Howes, B.K. Hamre, & R. C. Pianta (Eds.) *Effective Early Childhood Professional Development: Improving Teacher Practice and Child Outcomes.* Baltimore, MD: Paul H. Brookes.

Durksen, T. L., Way, J., Bobis, J., Anderson, J., Skilling, K., & Martin, A. J. (2017). Motivation and engagement in mathematics: a qualitative framework for teacher-student interactions. *Mathematics Education Research Journal, 29*(2), 163-181. doi: 10.1007/s13394-017-0199-1

Ehrenberg, R., Brewer, D., Gamoran, A., & Willms, J. (2001). Does Class Size Matter? *Scientific American, 285*(5), 78-85. Retrieved from http://www.jstor.org.login.ezproxy.library.ualberta.ca/stable/26059424

Emmer, E., & Stough, L. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist, 36*, 103–112. doi:10.1207/S15326985EP3602_5

Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., & Beechum, N. O. (2012). *Teaching adolescents to become learners. The role of noncognitive factors in shaping school performance: A critical literature review.* Chicago, IL: Consortium on Chicago School Research. Retrieved from https://files.eric.ed.gov/fulltext/ED542543.pdf

Field, A. (2016). *Exploring data: The beast of bias.* Retrieved from https://www.discoveringstatistics.com/repository/exploringdata.pdf

Freiberg, H. J., & Waxman, H. C. (1988). Alternative feedback approaches for improving student teachers' classroom instruction. *Journal of Teacher Education, 39*(4), 8-14. doi: 10.1177/002248718803900403

Friedman, S. (2014). *Teacher emotional intelligence and the quality of interactions with students* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Global. Rutgers University, New Jersey.

Gamlem, S. M., & Munthe, E. (2014). Mapping the quality of feedback to support students' learning in lower secondary classrooms. *Cambridge Journal of Education, 44*(1), 75-92. doi: 10.1080/0305764X.2013.855171

Gill, B., Shoji, M., Coen, T., & Place, K. (2016). The content, predictive power, and potential bias in five widely used teacher observation instruments. *Mathematica Policy Research.* Retrieved from Regional Educational Laboratory website at http://ies.ed.gov/ncee/edlabs.

Gitomer, D., Bell, C., Qi, Y., McCaffrey, D., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record, 116*(060304), 1-32.

Government of Alberta. (2010). *Inspiring action on education.* Retrieved from

https://www.oecd.org/site/eduilebanff/48763522.pdf

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal, 26*, 91–108. doi: 10.1111/j.1471-1842.2009.00848.x

Gregory, A., Allen, J. P., Mikami, A. Y., Hafen, C. A., & Pianta, R. C. (2014). Effects of a professional development program on behavioral engagement of students in middle and high school. *Psychology in the Schools, 51*(2), 143-163.

Gregory, A., Allen, J. P., Mikami, A. Y., Hafen, C. A., & Pianta, R. C. (2015). The promise of a teacher professional development program in reducing racial disparity in classroom exclusionary discipline. In D. Losen (Ed.) *Closing the school discipline gap: Equitable remedies for excessive exclusion*, 166-179. New York, NY: Teachers College Press.

Gregory, A., Hafen, C. A., Ruzek, E., Mikami, A. Y., Allen, J. P., & Pianta, R. C. (2016). Closing the racial discipline gap in classrooms by changing teacher practice. *School Psychology Review, 45*(2), 171-191.

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher, 43*(6), 293-303. doi: 10.3102/0013189X14544542

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship

between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education, 119*(3), 445-470. Retrieved from http://www.jstor.org/stable/10.1086/669901

Hafen, C. A., Allen, J. P., Mikami, A. Y., Gregory, A., Hamre, B., & Pianta, R. C. (2012). The pivotal role of adolescent autonomy in secondary school classrooms. *Journal of Youth and Adolescence, 41*(3), 245-255. doi: 10.1007/s10964-011-9739-2

Hafen, C. A., Hamre, B. K., Allen, J. P., Bell, C. A., Gitomer, D. H., & Pianta, R. C. (2015). Teaching through interactions in secondary school classrooms: Revisiting the factor structure and practical application of the Classroom Assessment Scoring System–Secondary. *The Journal of Early Adolescence, 35*(5-6), 651-680. doi: 10.1177/0272431614537117

Hafen, C. A., Ruzek, E. A., Gregory, A., Allen, J. P., & Mikami, A. Y. (2015). Focusing on teacher–student interactions eliminates the negative impact of students' disruptive behavior on teacher perceptions. *International Journal of Behavioral Development, 39*(5), 426-431. doi: 10.1177/0165025415579455

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials In Quantitative Methods For Psychology, 8*(1), 23. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3402032/

Halpin, P. F., & Kieffer, M. J. (2015). Describing profiles of instructional practice: a new approach to analyzing classroom observation data. *Educational Researcher, 44*(5), 263-277. doi: 10.3102/0013189X15590804

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development, 76*(5), 949-967. doi: 0009-3920/2005/7605-0001

Harms, T., & Clifford, R. (1998). *The Early Childhood Environment Rating Scale*. New York, NY: Teachers College Press.

Hattie, J. (2003). Teachers make a difference: What is the research evidence? *Distinguishing Expert Teachers from Novice and Experienced Teachers*. Camberwell: Australian Council for Educational Research. Retrieved from http://start.sd34.bc.ca/innovation/wp-content/uploads/2014/01/Hattie_TeachersMakeADifference.pdf

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. doi: 10.1080/10705519909540118

Huang, F. L. (2017). *Conducting multilevel confirmatory factor analysis using R.* doi: 10.13140/RG.2.2.12391.34724

Jamil, F. M., Downer, J. T., & Pianta, R. C. (2012). Association of pre-service teachers'

performance, personality, and beliefs with teacher self-efficacy at program completion. *Teacher Education Quarterly, 39*(4), 119-138. Retrieved from http://files.eric.ed.gov/fulltext/EJ1001446.pdf

Joe, J. N. McClellan, C. A., & Holtzman, S. L. (2014). Scoring design decisions: Reliability and the length and focus of classroom observations. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidelines from the Measures of Effective Teaching project* (pp. 415-443). San Francisco: Josey-Bass. Retrieved from http://k12education.gatesfoundation.org/wp-content/uploads/2015/11/Designing-Teacher-Evaluation-Systems_freePDF.pdf

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Research Paper. MET Project. Bill & Melinda Gates Foundation. Retrieved from http://files.eric.ed.gov/fulltext/ED540960.pdf

Klette, K., & Blikstad-Balas, M. (2018). Observation manuals as lenses to classroom teaching: Pitfalls and possibilities. *European Educational Research Journal, 17*(1), 129-146. doi: 10.1177/1474904117703228

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review, 47*, 180-195.

Kuncel, N. R., Credé, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. Review of Educational Research, 75(1), 63-82.

La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The classroom assessment scoring system: Findings from the prekindergarten year. *The Elementary School Journal, 104*(5). http://dx.doi.org/10.1086/499760

Lavigne, A. L., & Good, T. L. (2013). *Teacher and student evaluation: Moving beyond the failure of school reform.* New York: Routledge.

Luff, R., Byatt, D., & Martin, D. (2015). Review of the typology of research methods within the social sciences. National Centre for Research Methods Report. Retrieved from https://pdfs.semanticscholar.org/3c84/9e37fd4299eab0c5f1c0e98670f4b4ad99cf.pdf

Malmberg, L.-E., & Hagger, H. (2009). Changes in student teachers' agency beliefs during a teacher education year, and relationships with observed classroom quality, and day-to-day experiences. *British Journal of Educational Psychology, 79*, 677-694. doi: 10.1348/000709909X454814

Malmberg, L.-E., Hagger, H., Burn, K., Mutton, T., & Colls, H. (2010). Observed classroom quality during teacher education and two years of professional practice. *Journal of Educational Psychology, 102*(4), 916-932. doi: 10.1037/a0020920

Mashburn, A. J., Meyer, J. P., Allen, J. P., & Pianta, R. C. (2014). The effect of observation length and presentation order on the reliability and validity of an observational measure of teaching quality. *Educational and Psychological Measurement, 74*(3), 400-422. doi: 10.1177/0013164413515882

McCaffrey, D. F., Yuan, K., Savitsky, T. D., Lockwood, J. R., & Edelen, M. O. (2015). Uncovering multivariate structure in classroom observations in the presence of rater errors. *Educational Measurement: Issues and Practice, 34*(2), 34-46. doi: 10.1111/emip.12061

McCombs, B. L. (2013). Educational psychology and educational transformation. In I. B. Weiner, W. M. Reynolds, & G. E. Miller (Eds.), *Handbook of Psychology, Volume 7, Educational Psychology* (2nd ed., pp. 493-533). New Jersey: John Wiley & Sons, Inc.

Meyer, C. (2011). *Decatur, IL Eisenhower High School CLASS-S observations.* Research report. Naperville, IL: American Institutes for Research.

Meyers, C. V., Molefe, A., Brandt, W. C., Zhu, B., & Dhillon, S. (2016). Impact results of the eMINTS professional development validation study. *Educational Evaluation and Policy Analysis, 38*(3), 455-476. doi: 10.3102/0162373716638446

Mihaly, K., & McCaffrey, D. F. (2014). Grade-level variation in observational measures of teacher effectiveness. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher*

evaluation systems: New guidelines from the Measures of Effective Teaching project (pp. 9-49). San Francisco: Josey-Bass. Retrieved from http://k12education.gatesfoundation.org/wp-content/uploads/2015/11/Designing-Teacher-Evaluation-Systems_freePDF.pdf

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). A composite estimator of effective teaching. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from https://www.nbexcellence.org/cms_files/resources/Jan%202013%20%20A%20Composite%20Estimator%20of%20Effective%20Teaching%20Research%20Paper.pdf

Moher D, Liberati A, Tetzlaff J, Altman DG, & The PRISMA Group. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA statement. *PLoS Medicine, 6*(7): e1000097. doi: https://doi.org/10.1371/journal.pmed.1000097

National Association of School Psychologists. (2010). *Model for comprehensive and integrated school psychological services.* Retrieved from http://www.nasponline.org/standards/practice-model/

National Center for Teacher Effectiveness. (2012). *Mathematical Quality of Instruction.* Retrieved from http://isites.harvard.edu/icb/icb.do?keyword=mqi_training&tabgroupid=icb.tabgroup120173

Neergaard, L., & Smith, T. (2012). *Comparing beginning teachers' instructional quality growth on subject-specific and global measures.* Proceedings from the Society for Research on Educational Effectiveness. Retrieved from http://files.eric.ed.gov/fulltext/ED535664.pdf

The Organization for Economic Co-Operation and Development. (2015). *Programme for International Student Assessment (PISA).* Retrieved from http://www.oecd.org/pisa/

O'Sullivan, M. K., & Dallas, K. B. (2017). A collaborative approach to implementing 21st century skills in a high school senior research class. *Education Libraries, 33*(1), 3-9.

Ozer, E. J., & Douglas, L. (2015). Assessing the key processes of youth-led participatory research: Psychometric analysis and application of an observational rating scale. *Youth & Society, 47*(1), 29-50. doi: 10.1177/0044118X12468011

Pakarinen, E., Kiuru, N., Lerkkanen, M. K., Poikkeus, A. M., Siekkinen, M., & Nurmi, J. E. (2010). Classroom organization and teacher stress predict learning motivation in kindergarten children. *European Journal of Psychology of Education, 25*(3), 281-300. doi: 10.1007/s10212-010-0025-6

Park, Y. S., Chen, J., & Holtzman, S. L. (2014). Evaluating efforts to minimize rater bias in scoring classroom observations. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidelines from the Measures of Effective Teaching project* (pp. 383-414). San Francisco: Josey-Bass. Retrieved from

http://k12education.gatesfoundation.org/wp-content/uploads/2015/11/Designing-Teacher-Evaluation-Systems_freePDF.pdf

Patton, M. Q. (2011). *Developmental evaluation: Applying complexity concepts to enhance innovation and use.* New York: Guildford Press.

Peters, M. D. J., Godfrey, C. M., Khalil, H., McInerney, P., Parker, D., & Soares, C. B. (2015). Guidance for conducting systematic scoping reviews. *International Journal Of Evidence-Based Healthcare, 13*(3), 141-146. doi: 10.1097/XEB.0000000000000050

Pianta, R. C. (2006). Classroom management and relationships between children and teachers: implications for research and practice. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 685-709). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F. J. (2008). Classroom effects on children's achievement trajectories in elementary school. American *Educational Research Journal, 45*(2), 365-397. doi: 10.3102/0002831207308230

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109-119.

Pianta, R. C., Hamre, B. K., & Mintz, S. (2012). *Classroom Assessment Scoring System:*

Secondary manual. Charlottesville, VA: Teachstone.

Pianta, R. C., La Paro, K., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS).* Baltimore, MD: Paul H. Brookes.

Preston, C., Goldring, E., Guthrie, J. E., Ramsey, R., & Huff, J. (2017). Conceptualizing essential components of effective high schools. *Leadership and Policy in Schools, 16*(4), 525-562. doi: 10.1080/15700763.2016.1205198

Puentadura, R. (2010). *Substitution, Adaptation, Modification, and Redefinition Model.* Retrieved from www.hippasus.com/rrpweblog/archives/2012/08/23/SAMR_BackgroundExemplars.pdf

Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning. *Educational Researcher, 29*(1), 4-15. Retrieved from http://www.jstor.org/stable/1176586

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Ripski, M. B., LoCasale-Crouch, J., & Decker, L. (2011). Pre-service teachers: Dispositional traits, emotional states, and quality of teacher-student interactions. *Teacher Education Quarterly, 38*(2), 77-96. Retrieved from http://files.eric.ed.gov/fulltext/EJ926861.pdf

RStudio Team (2018). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA.

Retrieved from http://www.rstudio.com/.

Rose, D.H. & Meyer, A. (2006). *A Practical Reader in Universal Design for Learning.* Cambridge, MA: Harvard Education Press.

Rosseel, Y. (2017). lavaan: An R package for structural equation modeling. Journal of Statistical Software, 48(2), 1-36. Retrieved from http://www.jstatsoft.org/v48/i02/

Rutledge, S., Cohen-Vogel, L., & Osborne-Lampkin, L. (2012). *Identifying the characteristics of effective high schools: Report from year one of the National Center on Scaling Up Effective Schools.* Research report. National Center on Scaling Up Effective Schools (NJ1). Retrieved from http://files.eric.ed.gov/fulltext/ED538011.pdf

Ruzek, E. A., Hafen, C. A., Allen, J. P., Gregory, A., Mikami, A. Y., & Pianta, R. C. (2016). How teacher emotional support motivates students: The mediating roles of perceived peer relatedness, autonomy support, and competence. *Learning and Instruction, 42*, 95-103. doi: 10.1016/j.learninstruc.2016.01.004

Ruzek, E. A., Hafen, C. A., Hamre, B. K., & Pianta, R. C. (2014). Combining classroom observations and value added for the evaluation and professional development of teachers. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidelines from the Measures of Effective Teaching project* (pp. 205-233). San Francisco: Josey-Bass. Retrieved from http://k12education.gatesfoundation.org/wp-

content/uploads/2015/11/Designing-Teacher-Evaluation-Systems_freePDF.pdf

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68. doi: 10.1037110003-066X.55.1.68

Sanders, W. L., Wright, S. P., & Horn, S. P. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. Journal of Personnel Evaluation in Education, 11(1), 57-67. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.581.9316&rep=rep1&type=pdf

Sandstrom, H. (2012). The characteristics and quality of pre-school education in Spain. *International Journal of Early Years Education, 20*(2), 130-158. doi: 10.1080/09669760.2012.715012

Savitsky, T. D. & McCaffrey, D. F. (2014). Bayesian hierarchical multivariate formulation with factor analysis for nested ordinal data. *Psychometrika, 79*(2), 275-302. doi: 10.1007/S11336-013-9339-Z

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods Of Psychological Research Online, 8*(2), 23-74. Retrieved from http://citeseerx.ist.psu.edu.login.ezproxy.library.ualberta.ca/viewdoc/download?doi=10.1.1.

509.4258&rep=rep1&type=pdf

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454-499. doi: 10.3102/0034654307310317

Shyamalan, M. N. (2013). *I Got Schooled: The Unlikely Story of How a Moonlighting Movie Maker Learned the Five Keys to Closing America's Education Gap*. New York: Simon and Schuster.

Shell, M. D., Gazelle, H., & Faldowski, R. A. (2014). Anxious solitude and the middle school transition: A diathesis x stress model of peer exclusion and victimization trajectories. *Developmental Psychology, 50*(5), 1569-1583. doi: 10.1037/a0035528

Shernoff, D. J., Ruzek, E. A., & Sinha, S. (2017). The influence of the high school classroom environment on learning as mediated by student engagement. *School Psychology International, 38*(2), 201-218. doi: 10.1177/0143034316666413

Smith, T. M., Cannata, M., & Haynes, K. T. (2016). Reconciling data from different sources: Practical realities of using mixed methods to identify effective high school practices. *Teachers College Record, 118*(7), 1-34.

Smith, T., Preston, C., Haynes, K., & Neergaard Booker, L. (2015). Understanding differences in instructional quality between high and lower value-added schools in a large urban

district. *Teachers College Record, 117*(11), 1-38.

Smith, V., Hayes, S., Labonté, C., & Vargas, F. (2016). *Flexible pathways to success: Technology to design for diversity.* Retrieved from the Government of Alberta website: https://open.alberta.ca/dataset/029047d9-3a06-4b18-9c94-40d0c3937639/resource/e67821ad-ce9a-4f73-9248-b9710ebd6cf2/download/uofa-flexible-pathways-finalreport-july26-2016-online.pdf-

Spearman, J., & Watt, H. M. G. (2013). Perception shapes experience: The influence of actual and perceived classroom environment dimensions on girls' motivations for science. *Learning Environments Research, 16*, 217-238. doi: 10.1007/s10984-013-9129-7

Stanford University. (2013). *PLATO: Protocol for Language Arts Teaching Observations.* Retrieved from http://platorubric.stanford.edu/index.htmlhttp://platorubric.stanford.edu/index.html

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education, 62*(4), 339-355. doi: 10.1177/0022487111404241

Stuhlman, M. W., Hamre, B. K., Downer, J. T., & Pianta, R. C. (2010). Why should we use classroom observation? *Centre for Advanced of Teaching and Learning,* University of Virginia. Retrieved from

https://curry.virginia.edu/uploads/resourceLibrary/CASTL_practioner_Part1_single.pdfU

NESCO. (2014). *Teaching and Learning: Achieving Quality for All.* Retrieved from http://unesdoc.unesco.org/images/0022/002256/225654e.pdf

Tondeur, J., Aesaert, K., Pynoo, B., Braak, J., Fraeyman, N., & Erstad, O. (2017). Developing a validated instrument to measure pre-service teachers' ICT competencies: Meeting the demands of the 21st century. *British Journal of Educational Technology, 48*(2), 462-472. Retrieved from https://biblio.ugent.be/publication/7036378/file/7036560

Van Laar, E., van Deursen, A. J., van Dijk, J. A., & de Haan, J. (2017). The relation between 21st-century skills and digital skills: A systematic literature review. *Computers in Human Behavior, 72*, 577-588. doi: 10.1016/j.chb.2017.03.010

Vaske, J. J., Beaman, J., & Sponarski, C. C. (2017). Rethinking internal consistency in Cronbach's alpha. *Leisure Sciences, 39*(2), 163-173. doi: 10.1080/01490400.2015.1127189

Vermette, D., Foote, D., Bird, C., Mesibov, D., Harris-Ewing, D., & Battaglia, D. (2001). Understanding constructivism(s): a primer for parents and school board members. *Education, 122*, 87–93.

Virtanen, T. E., Lerkkanen, M.-K., Poikkeus, A.-M., & Kuorelahti, M. (2015). The relationship between classroom quality and students' engagement in secondary school. *Educational*

Psychology: An International Journal of Experimental Educational Psychology, 35(8), 963-983. doi: 10.1080/01443410.2013.822961

Virtanen, T. E., Pakarinen, E., Lerkkanen, M. K., Poikkeus, A. M., Siekkinen, M., & Nurmi, J. E. (2018). A validation study of Classroom Assessment Scoring System–Secondary in the Finnish school context. *The Journal of Early Adolescence, 38*(6), 849-880. doi: 10.1177/0272431617699944

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research 73*(1), 89-122. doi:10.3102/00346543073001089

Westergård, E., Ertesvåg, S. K., & Rafaelsen, F. (2018). A preliminary validity of the classroom assessment scoring system in Norwegian lower-secondary schools. *Scandinavian Journal of Educational Research,* 1-19. doi: 10.1080/00313831.2017.1415964

Wilson, H. K., Scult, M., Wilcher, M., Chudnofsky, R., Malloy, L., Drewel, E., ... & Denninger, J. W. (2015). Teacher-led relaxation response curriculum in an urban high school: impact on student behavioral health and classroom environment. *Advances in Mind-body Medicine, 29*(2), 6-14.

World Education Forum. (2015). Incheon Declaration Education 2030: Towards inclusive and equitable quality education and lifelong learning for all. Retrieved from http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/ED/ED/pdf/FinalVersion-

IncheonDeclaration.pdf

Wragg, T., & Wragg, E. C. (2012). *An Introduction to Classroom Observation.* New York, NY: Routledge.

Wubbels, T. (2014). Teacher-student relationships in class: A future for research [Abstract]. *Pedagogische Studien, 91*(5), 352-363.

Yoder, N. P. (2013). *Linking social-emotional learning to teaching practices in an urban context: A mixed-methods study* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Global. (UMI 3610176).

Yuan, K., McCaffrey, D. F., & Savitsky, T. D. (2013). *Analyzing the factorial structure of the Classroom Assessment Scoring System-Secondary using a Bayesian hierarchical multivariate ordinal model.* Paper presented at the Society for Research on Educational Effectiveness, Washington, DC. Retrieved from http://files.eric.ed.gov/fulltext/ED563074.pdf

Zohar, A., & Dori, Y. (2003). Higher order thinking skills and low achieving students: Are they mutually exclusive? *Journal of the Learning Sciences, 12*, 145–181. doi:10.1207/S15327809JLS1202_1

Appendix A

Research Ethics Approval Letter

Notification of Approval

Date:	July 26, 2016
Study ID:	Pro00066357
Principal Investigator:	Stephanie Hayes
Study Supervisor:	Veronica Smith
Study Title:	Observing Teacher-Student Interactions Using the Classroom Assessment Scoring System – Secondary (CLASS-S) in Alberta
Approval Expiry Date:	Tuesday, July 25, 2017
Sponsor/Funding Agency:	Alberta Education 4801

	Project ID	Project Title	Speed Code	Other Information
RSO-Managed Funding:	RES0017912	Research project to gain a deep understanding of promising practices, implementation strategies and educational benefits/results of technology use within inclusive Junior High/Middle School settings		

Thank you for submitting the above study to the Research Ethics Board 2. Your application has been reviewed and approved on behalf of the committee.

A renewal report must be submitted next year prior to the expiry of this approval if your study still requires ethics approval. If you do not renew on or before the renewal expiry date, you will have to re-submit an ethics application.

Approval by the Research Ethics Board does not encompass authorization to access the staff, students, facilities or resources of local institutions for the purposes of the research.

Sincerely,

Stanley Varnhagen, PhD
Chair, Research Ethics Board 2

Note: This correspondence includes an electronic signature (validation and approval via an online system).