Assessing Learning Skills and Work Habits: What Do Report Card Data Tell Us?

Stefan Denis Merchant
Queen's University

Don Klinger
University of Waikato

John Kirby
Queen's University

Abstract

Many school systems ask teachers to assess and report upon aspects of student performance beyond academic achievement. In Ontario, K–12 teachers assess a common set of six Learning Skills and Work Habits. How well teachers are able to undertake these assessments is not well studied. This study examines Grade 9 and 12 report card data from two districts in Ontario, Canada to determine to what extent different learning skills are assessed independently of each other, and to what extent they are associated with teacher-awarded academic achievement and achievement on a standardized Grade 9 mathematics examination delivered by Ontario's Educational Quality and Accountability Office (EQAO). Results indicate that the set of six Learning Skills and Work Habits

Canadian Journal of Education / Revue canadienne de l'éducation 45:3 (2022)

www.cje-rce.ca

are assessed as a unitary construct. Grades on these skills have higher correlations with teacher-awarded grades than with standardized test scores. Finally, gender differences in both academic achievement and achievement on the set of skills are investigated.

Keywords: learning skills, report cards, classroom assessment, self-regulation, 21st century skills

Résumé

Plusieurs systèmes scolaires s'attendent des enseignants qu'ils évaluent et rendent compte des différents aspects de la performance des élèves, au-delà de la réussite scolaire. En Ontario, tous les enseignants (de la maternelle à la 12e année) évaluent un ensemble commun de six habiletés d'apprentissage et habitudes de travail. La capacité des enseignants à effectuer ces évaluations n'est cependant pas bien étudiée. Cette étude examine donc les données des bulletins scolaires de 9e et de 12e année de deux commissions scolaires d'Ontario, afin de déterminer dans quelle mesure les différentes compétences d'apprentissage sont 1) évaluées indépendamment les unes des autres, 2) associées à la réussite scolaire et 3) corrélées avec les résultats d'un examen de mathématiques standardisé. Les résultats indiquent que l'ensemble des six habiletés d'apprentissage et habitudes de travail est évalué comme un construit unitaire. L'examen des notes sur ces habiletés d'apprentissage et habitudes de travail révèle des corrélations plus élevées avec les notes attribuées par les enseignants qu'avec les résultats de l'examen standardisé. Enfin, les différences entre les sexes dans la réussite scolaire et la réussite sur l'ensemble des habiletés d'apprentissage et habitudes de travail ont été étudiées.

Mots-clés : habiletés d'apprentissage, bulletins scolaires, évaluations en salle de classe, auto-régulation, compétences du XXIe siècle

Introduction

Teachers are commonly expected to assess elements of student performance beyond academic achievement. These assessments include constructs such as effort, participation, and collaboration. Examples of such expectations for student evaluation can be found in many countries, including Canada, Northern Ireland, and Singapore (Merchant et al., 2018; Northern Ireland Ministry of Education, 2007; Singapore Ministry of Education, 2021). Similar expectations are included in international education systems such as the International Baccalaureate program, and the International Primary Curriculum (International Baccalaureate Organization, 2009; Fieldwork Education, n.d.). Educational organizations in the United States have also acknowledged the need for teachers to evaluate educational outcomes beyond achievement. The Association for Supervision and Curriculum Development (ASCD) argued for a new learning compact, one that focuses not just on academics, but also on developing other factors such as empathy, curiosity, creativity, self-discipline, and social competence (ASCD, 2007). The National Education Association (NEA; 2012) initiated a discussion about how to develop and assess critical thinking, communication, collaboration, and creativity (collectively known as the "4 Cs") in American public schools. Thus, there appears to be widespread agreement that it is desirable for schools to develop and assess students′ educational outcomes that reflect skills beyond subject area achievement.

While there is broad consensus that classroom teachers should assess learning skills and other competencies, there is no consensus on an appropriate umbrella term for such skills, or of the components that such skills should encompass (Duckworth & Schulze, 2009). Within Canada, a broad range of terms is used, such as "Learner Profile," "Learning Behaviours," and "Competencies" (Merchant et al., 2018). Ontario currently uses "Learning Skills and Work Habits" (LSWH) but has investigated using the term "21st Century Competencies" (Ontario Ministry of Education, 2016). In this article, we will use the term "learning skills" to refer to general skills that support students' learning in schools, and Learning Skills and Work Habits (LSWH) to refer to the six specific skills included on Ontario report cards.

A strong rationale supports including learning skills as part of teachers′ assessment and evaluation of their students. Firstly, a large research base connects these skills with improved learning (e.g., Farrington et al., 2012; Jacob, 2002; Muenks et al., 2017;

Zimmerman, 1990). Skills such as metacognition, self-regulation, and self-efficacy are significantly and positively correlated with learning (Ivcevic & Brackett, 2014; Kleitman & Costa, 2014; Zimmerman & Kitsantas, 2014; Zuffianò et al., 2013). In addition, economic research has found that traits such conscientiousness and agreeableness (related to LSWH such as responsibility and collaboration) are associated with better long-term outcomes such as higher employment income, relationship stability, health, lower criminality, and lower drug use (see for example: Almlund et al., 2011; Borghans et al., 2008). Finally, employers have identified skills such as initiative, planning and organizing time, and an independent work ethic as vital to workplace performance and are demanding that students develop these skills in schools (Casner-Lotto & Barrington, 2006; Conference Board of Canada, 2015; Levin, 2012).

There is a further argument that assessing and reporting separately on achievement and non-achievement factors yields a more complete picture of student performance. It is known that teachers' achievement grades are not pure measures of achievement (Cross & Frary, 1999). Instead, they reflect a mix of constructs that includes not only achievement, but factors such as effort, focus, and improvement (McMillan, 2001). Some classroom assessment experts (e.g., Guskey et al., 2011) have argued that by reporting learning skills separately from achievement we obtain a more complete picture of student performance in the classroom. However, this argument is predicated on the assumption that teachers can and do grade achievement and non-achievement factors separately. If, for example, the grades teachers award for learning skills are based upon achievement, then the additional information provided by these grades may be minimal.

While the rationale to assess and evaluate learning skills is strong, there is an open question as to how well teachers can do this. This may be especially true of secondary school teachers, who spend less time with their students than elementary school teachers. Many of these skills are inconsistently defined in the research literature and having a precise definition of a measurable construct is critical if assessment and grading are to be reliable and to result in valid interpretations of students′ skills (Bass, 2005). In addition, many teachers struggle with assessment and grading (e.g., Cizek, 1996; DeLuca & Bellara, 2013). Further, there is evidence that assessment education in teacher preparation programs focuses on assessment of subject matter knowledge and skills, and not on more general skills such as collaboration or perseverance (Poth, 2012). Our own scan of assessment courses offered by teacher education programs in Ontario found only one program

(out of 13) that offered a course that addresses how to assess LSWH. These facts call into question whether teachers can effectively assess their students′ learning skills in their classrooms, and if teachers′ grading of learning skills yields useful, actionable information.

The Ontario Context

All Ontario K–12 teachers must assess, grade, and report upon a set of six LSWH. The six LSWH are: collaboration, initiative, independent work, organization, responsibility, and self-regulation, and are considered to be "an integral part of students' learning" (Ontario Ministry of Education, 2010, p. 10). A 4-point scale is used for reporting performance on the LSWH, with teachers selecting from "Excellent," "Good," "Satisfactory," and "Needs Improvement." Teachers report on the LSWH halfway through a course, and at the end.

The Ontario Ministry of Education (2010) does not define these six skills. Rather, they provide teachers with examples of observable classroom behaviours that may serve as indicators of the skills. For example, behaviours associated with "organization" include "devises and follows a plan and process for completing work and tasks," and "establishes priorities and manages time to complete tasks and achieve goals" (p. 10). While these descriptors are very reasonable for organization, they also very closely match descriptions of self-regulated learning found in the literature (Hadwin & Winne, 2012; Zimmerman, 2013). "Self-regulation" has associated behaviours of "assesses and reflects critically on own strengths, needs, and interests" and "perseveres and makes an effort when responding to challenges" (Ontario Ministry of Education, 2010, p. 10). These behaviours also match definitions of grit and metacognition (Duckworth & Gross, 2014; Flavell, 1979).

The example behaviours described here illustrate the challenge in precisely defining and distinguishing learning skills such as "organization," or "initiative." Adding to the challenge, these skills are context dependent—organization may look different in an English classroom than it does in a physical education classroom. Further, these skills are not independent (Diamond, 2013; Stecher & Hamilton, 2014). Organization requires self-regulation, and a sense of responsibility toward the group is necessary for effective collaboration. The complex, context-dependent, and interrelated nature of these skills may explain why researchers and policy makers have struggled to provide consistent

and distinct definitions of learning skills (Duckworth & Schulze, 2009; Farrington et al., 2012). Of importance for our work, if those responsible for deeply understanding these skills struggle to define them precisely, a legitimate question arises as to how teachers define these skills. Imprecise definitions lead to poor construct validity and interpretations in assessment (Watson & Emery, 2010). If parents, students, and other stakeholders are to interpret the meaning of the LSWH accurately it is necessary that we understand how teachers define the LSWH. Without this understanding, it is difficult to see how the LSWH grades may be used for any type of educational purpose.

The difficulties of defining LSWH are just the first hurdle with respect to assessing LSWH. Even if teachers are devising precise definitions of the LSWH, they must also create good measures of the LSWH. Measuring a construct becomes more difficult when the construct is multi-dimensional and varies depending on context. We can use self-regulation (one of the Ontario LSWH) as an example. Not only are there a variety of definitions of self-regulation (e.g., Baumeister & Vohs, 2007; Dinsmore et al., 2008; Schunk, 2008; Zimmerman, 2000), but student self-regulation is known to vary depending on motivation (Pintrich & De Groot, 1990). Self-regulation has been measured using a broad variety of instruments. These include: self-report questionnaires such as the Motivated Strategies for Learning Questionnaire (Dinsmore et al., 2008), standardized observation protocols (Zimmerman & Kitsantas, 2014), interviews (Zimmerman & Martinez-Pons, 1988), specific tasks (Galla et al., 2014), think-alouds (Greene et al., 2011), and trace data (Winne & Perry, 2000). These tools and methods have various advantages and disadvantages, but one common issue with these assessments of self-regulation in the research literature is that they tended to be conducted once, or over a short period of time. This is not reflective of the classroom environment.

Assessments occurring at a single point in time are not able to measure the "habit" portion of LSWH. A student who demonstrates strong LSWH on a single task may or may not consistently demonstrate such behaviours in the classroom. Accurate and meaningful assessment of the LSWH requires sustained interaction with the students. Hence, it is reasonable to consider teachers as ideally positioned to make these assessments (Zimmerman & Martinez-Pons, 1988). Teachers have daily interactions with their students over an extended period of time and can potentially offer a context-rich perspective that is missing from assessments of a single event or at a single point in time such as a task-based measure, self-report questionnaire, or think-aloud process. Teachers

are also exposed to a range of self-regulatory styles and capabilities among their students, enabling them to have a sense of how a student's self-regulation compares with norms of the current classroom and the teacher's prior classrooms (Wigelsworth et al., 2010). Finally, teachers can assess self-regulation using a variety of tools. Nothing prevents teachers from using questionnaires or interview protocols with their students. These types of data could then be supplemented with day-to-day observational data and student self-assessments, such as reflections or journals, allowing for the collection of a rich data set from which to make a judgement about a student's ability to self-regulate.

While the potential for teachers to be good assessors of LSWH may be high, certain realities cannot be ignored. For instance, many aspects of LSWH are internal to the student, and difficult to observe directly. The Ontario Ministry of Education (2010) describes one aspect of self-regulation as "reflects critically on own strengths" (p. 11), but how does a teacher observe this? Student self-assessment may provide one useful tool, but the reality is that latent thought processes are not directly observable, and therefore difficult for classroom teachers to assess (Lai, 2011). Further, there is strong evidence that teachers' grading practices are inconsistent and idiosyncratic (Bowers, 2011; Brown, 2011; Howley et al., 2000; Lekholm & Cliffordson, 2008). If teachers experience challenges with assessing subject area achievement, is it reasonable to expect them to be competent at assessing LSWH? Previous research has highlighted that teachers report struggling to assess simpler aspects of LSWH such as student effort and participation (Linn & Miller, 2005; Miller et al., 2006). For instance, it is easy to conflate effort with achievement, and participation can take many forms, some of which are not observable.

One study that directly addressed how teachers assess non-achievement constructs was conducted by Ferrito (2015). He found that teachers struggle to assess different "Personal and Social Development" items independently. In his study of the report cards of 113 Grade 4 students in New Jersey, Ferrito found that teachers' ratings of seven such items were best described using a one-factor model. This model accounted for 75% of the variance in the ratings, and all constructs loaded at 0.84 and above onto the single factor. Examples of the "Personal and Social Development" constructs include "Is able to follow classroom directions," "Is able to follow rules," and "Is able to use Listening Position" (p. 74). A surface inspection of these constructs indicates they may be all related to compliance, and so a one-dimensional factor structure is perhaps not surprising.

Ferrito's results may also rise from a halo effect in teachers' ratings of "Personal and Social Development." In his discussion of the halo effect, Thorndike (1920) stated that "even a very capable foreman, employer, teacher, or department head is unable to treat an individual as a compound of separate qualities and to assign a magnitude to each of these in independence of the others" (p. 28). Thus, it is possible that while teachers may be required to assess different learning skills as independent constructs, they are incapable of doing so. There is some research to support this hypothesis. Babad and colleagues (1982) found that physical education teachers' ratings of students' potential were impacted by irrelevant factors such as socio-economic status and physical attractiveness. Duckworth and Yeager (2015) suggested that "Teachers' ratings of students' specific qualities can also be colored by their top-down, global evaluations" (p. 241). The halo effect appears in many different rating contexts such as student ratings of instructors (Keeley et al., 2013), supervisor ratings of medical residents (McGill et al., 2011), and ratings of students' academic engagement (Briesch et al., 2010). Given these findings, there remains an open question as to how well teachers can assess different learning skills as independent constructs.

Our own qualitative work in Ontario revealed that secondary teachers struggled to articulate how the six LSWH were distinct—except for collaboration (Merchant, 2016). During interviews, teachers reported collaboration was the easiest LSWH to assess because it is visible in the classroom. Further, collaboration was the only LSWH for which teachers had specific assignments or tasks that served as assessments. Based on these data, we concluded that Ontario secondary teachers hold a two-dimensional view of the LSWH, with collaboration forming one dimension, and the other five LSWH coalescing into a second dimension. This view is consistent with the perspective that 21st century skills can be divided into interpersonal and intrapersonal skills (Pellegrino & Hilton, 2013; Stecher & Hamilton, 2014).

Description of the Study

The paucity of research on how teachers define, assess, and report upon learning skills is surprising as there have long been calls for research to be done in this area (e.g., McMillan & Workman, 1998; Stecher & Hamilton, 2014). Our research not only acknowledges the need for such research, but also more deeply explores the issue than was found in pre-

vious studies. In doing so, we hope to illuminate teacher practices in this important, but poorly understood area of classroom assessment. Our efforts to understand how secondary teachers define, assess, and report upon the six LSWH were supported by a large set of secondary school report card data. Three specific research questions guided our work:

1. To what extent do Ontario secondary school teachers assess the six LSWH independently of each other?
2. To what extent are Ontario secondary school teachers able to assess the six LSWH independently of academic achievement?
3. What gender differences exist in patterns of grades on the LSWH?

Report card data were obtained from two school districts in Ontario, Canada. The dataset included all Grade 9 and 12 students within each district. For each student, we received the course code (e.g., Grade 9 English Applied Level), the final grade, the final LSWH grades, the number of absences, and the number of lates. District 1 raw data consisted of 57,230 sets of grades, but 982 of those were missing the LSWH component. We could see no obvious patterns as to which courses or types of students did not have LSWH grades inputted but noted for students where the LSWH grade was missing, the achievement grade was also frequently missing. As an example, for District 1, of the 982 excluded data sets, 775 were also missing the achievement grade. For the excluded data sets that did have the achievement grade, the mean grade (M = 71.17, SD = 21.10) was not significantly different than that of the final sample (M = 71.99, SD = 16.85; t(30,816) = -0.70, p = 0.49). There was a significant difference in gender balance (c^2(31,794) = 7.82, $p < .01$) with fewer males (48.8%) in the excluded data than in the included data (53.3%). Unfortunately, we have no way of knowing why missing data is excluded, but because only 1.7% of the data were excluded we feel confident that our final sample for District 1 was representative of the population of that district.

District 2 raw data contained 26,024 sets of grades, but 1,004 of those were missing the LSWH components, and so were not usable. Patterns to the missing data were slightly different than for District 1. The gender balance was not significantly different (c^2(25,394) = 3.59, p = .06), but the mean achievement grade of the excluded sample (M = 70.26, SD = 16.57) was significantly lower than for the final sample (M = 76.44, SD = 13.67; t(25,252) = -13.15, $p < .001$). In this case, the excluded data accounts for 3.9% of the total data, and again we feel this number is small enough that our sample is likely representative of the population.

Ontario high school students typically take six to eight courses per year, so the actual number of students included in the sample is lower than the number of sets of grades. To comply with privacy requirements, data were made anonymous by the school districts so that no identifying information with regards to student, teacher, or school was available. This prevented us from conducting analyses that examined effects at the teacher or school level. So, while we know that District 1 has 15 secondary schools, and District 2 has four secondary schools, we do not know which grades came from which school, or which class. All statistical analyses were conducted using SPSS version 24.0, except for the confirmatory factory analysis, which was conducted using the student edition of LISREL 9.2.

To answer our first research question (RQ1), an exploratory factor analysis was conducted on the LSWH report card data from one school district. The data from the second school district were then analyzed using a confirmatory factor analysis to determine if the factor structure remained constant across districts. Based upon findings from our prior qualitative study (Merchant, 2016), we hypothesized that a two-dimensional factor structure would emerge. We expected collaboration to be a distinct factor, and that the other five LSWH would form a second factor.

Our second research question (RQ2) was answered using correlational analyses between LSWH grades, teacher awarded subject grades and scores on a standardized mathematics examination. Students in Ontario receive only two standardized tests during high school, and one of those tests is a minimum competency literacy test graded on a pass/fail basis, and therefore not suitable for correlational analyses. The other test is a Grade 9 mathematics examination. Hence, the data used to answer this research question were restricted to Grade 9 mathematics only. Based on Steiger's work (1980), we used Fisher's r to z transformation to test whether the correlation coefficients were significantly different. We hypothesized that LSWH grades would show a stronger correlation with teacher-awarded grades than with standardized test scores. This hypothesis was based upon earlier findings that teachers include constructs such as effort, participation, and attendance in their achievement grades, whereas these constructs are absent from standardized testing results (Brookhart, 1993; Cross & Frary, 1999; Russell & Austin, 2010). Hierarchical multiple regression was used to examine whether LSWH grades could predict standardized mathematics examination scores beyond the teacher-awarded grade.

Our third research question (RQ3) was answered by comparing means using *t*-tests and calculating effect sizes. A further analysis was conducted separately for Grade 9 physical education, as this is the only course in which students are separated by gender. By exploring if gender differences in grades remain constant in single gender courses, it is possible to illuminate the extent to which LSWH grades are norm referenced. Based upon findings by Duckworth and Seligman (2006), we hypothesized that girls would be awarded higher LSWH grades than boys.

Results

The first RQ was answered using exploratory factor analysis for the LSWH ratings. Ratings were recoded to a numerical scale so that the top point of the scale (excellent) equated to a 4, and the bottom point (needs improvement) equated to 1. The first analysis focused on the Grade 9 report card data from District 1 only (n = 31,087). The Kaiser-Meyer-Olkin measure of sampling adequacy (KMO = 0.945) demonstrated that patterns of correlation were compact, meaning exploratory factor analysis would likely yield interpretable results. Exploratory factor analysis (EFA) was conducted using a maximum likelihood algorithm within SPSS (ver. 24.0). A scree plot was used to determine the dimensionality of the data set. The EFA revealed a unidimensional factor structure with the single factor accounting for 82.1% of the variance. The factor loadings are provided in Table 1. The EFA was repeated for the Grade 12 data (n = 22,854) within the same district, and the results were nearly identical. The single factor accounted for 82.6% of the variance (see Table 1 for loadings).

Table 1

Factor Loadings for Grades 9 and 12 in District 1

	District 1	
LSWH	Grade 9 Factor Loading	Grade 12 Factor Loading
Collaboration	.85	.84
Independent Work	.91	.91
Initiative	.92	.93
Organization	.90	.91
Responsibility	.93	.93
Self-regulation	.92	.92

To determine if the factor structure was consistent across contexts, the District 1 data were reanalyzed separately for each course. For statistical purposes, only courses where the district-wide enrolment was greater than 100 were included. The factor structure remained consistent across all courses in both Grades 9 and 12. Of interest, collaboration was always the lowest loading LSWH, and the remaining five LSWH were tightly clustered in terms of factor loading, although the ordering was not identical across courses. The amount of variance accounted for by the single factor tended to be lowest in mathematics and science courses. As an example, Grade 12 biology exhibited the lowest amount of variance accounted for by the single factor, with the single factor accounting for 73.43% of the variance. At the other extreme, a single factor accounted for 89.23% of the variance in Grade 9 music. The EFA was further repeated separately for each gender. Once again, the factor structure remained consistent, with a single factor accounting for 80.17% of the variance for boys, and 82.79% of the variance for girls. Collaboration retained the lowest factor loading (0.83 for boys, and 0.86 for girls), and the other five LSWH were all above 0.90, except for boys' responsibility, which had a factor loading of 0.88.

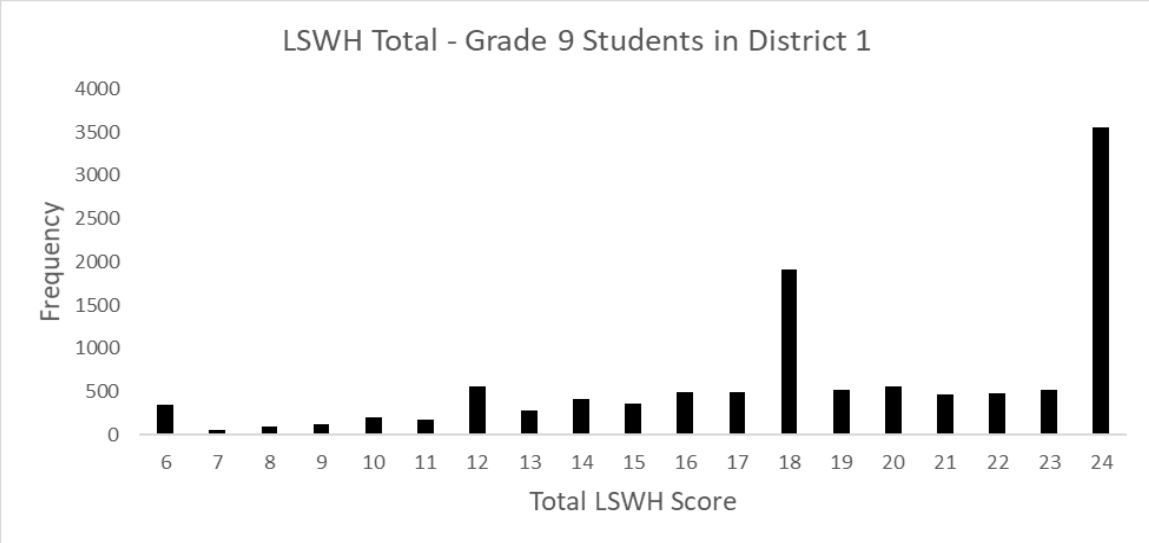
With the EFA giving such strong evidence for a one-dimensional factor structure, it is not surprising that the CFA confirmed that a one-factor model was appropriate for these data. The CFA was completed using the Grade 9 data from District 2, and most model fit parameters were very good. The single factor model yielded CFI = 0.99, with a standardized root mean square residual (SRMSR) of 0.01. In contrast, the finding that RMSEA = 0.088, CI [0.083, 0.092] was surprising. We repeated the analyses treating the LSWH grades as ordered data. This change made little difference to the numerical results, and no difference to the general conclusions. Treating the data as ordered, the one-factor CFA gave fit indices of CFI = 0.999, SRMSR = 0.009, and RMSEA = 0.105, CI [0.102, 0.109]. The high RMSEA values are not ideal, but there is evidence that models with few variables have an inflated RMSEA value (Kenny & McCoach, 2003), so we do not take the high RMSEA as evidence of a poorly fitting model. We would have liked to test other models, but the only other model with a theoretical justification was a two-dimensional model with collaboration as the second dimension. Because collaboration has only a single measure (the collaboration grade), if we allow collaboration to correlate freely with the other dimension, this two-dimensional model is mathematically identical to a single factor model. The correlation between collaboration and the other dimension will be identical to the factor loading of correlation in the single factor model. Thus, we gain no new

information in testing this model. Instead, as an indicator of how well a two-dimensional model would fit the data, we used EFA and forced a two-factor model onto the data. The variance accounted for increased by 0.79% in District 1, and 1.71% in District 2. None of the LSWH loaded onto the second factor.

Another way of visualizing the data is through a histogram of the total LSWH scores. For each set of LSWH grades, a total score was calculated, by summing the six individual LSWH grades. There are spikes at total LSWH scores of 6, 12, 18, and 24 (Figure 1). The histogram further reveals a negative skew to the Grade 9 data (skewness = -0.567 for District 1, and -0.888 for District 2). This skewness is due to a pronounced ceiling effect that is occurring with the LSWH grades. It was found that 45% of students in District 1 and 53% of students in District 2 received the same grade for all six LSWH.

Figure 1

Histogram of LSWH grades awarded



The total LSWH score was also used to address the second research question. Since teachers' ratings of the six LSWH were unidimensional, summing the six LSWH ratings created a good measure of student performance with respect to LSWH. Students′ total LSWH scores were then correlated to both the teacher-awarded final grade in the course, and to the score the student received on a province-wide standardized mathematics exa-

mination. As the only applicable standardized test given to high school students in Ontario is for Grade 9 mathematics, the analyses were restricted to this course. High school students in Ontario are streamed into either applied or academic mathematics courses. According to the Ontario Ministry of Education (2005), applied mathematics focuses "on the essential concepts of a subject" to "develop students' knowledge and skills through practical applications and concrete examples" (p. 6). In practice, students are often streamed into applied mathematics when teachers feel the student would struggle with the demands of the academic mathematics course. For these analyses, the two courses were separated. Results from the two districts showed that the total LSWH scores were more strongly correlated with teacher-awarded grades, than with standardized test scores (see Table 2). Using Fisher's *r* to *z* transformation, it was determined that for both districts and both streams, the differences in correlation coefficients were significant ($p < 0.001$).

Table 2

Pearson Correlation Coefficients between Academic Achievement and LSWH Total

	District 1		District 2	
	Applied Mathematics ($N = 1{,}326$)	Academic Mathematics ($N = 2{,}122$)	Applied Mathematics ($N = 242$)	Academic Mathematics ($N = 494$)
Teacher awarded final grade	0.79	0.78	0.80	0.77
Standardized mathematics examination score	0.52	0.53	0.45	0.56
z-value (Fisher's r)	12.47**	15.02**	6.86**	6.23**

** $p < 0.001$

To further investigate the relationship between LSWH and achievement, a hierarchical multiple regression was conducted to determine if LSWH contributed to the prediction of the standardized mathematics examination score. Data from District 2 were used as they contained information about students' attendance. Gender, lates, and absences were entered into the first step of the regression, the teacher-awarded grade into the second step, and the LSWH total into the third step (Table 3). The final model accounted for 63% of the variance in the standardized mathematics examination score. However, the addition of the third step added only 0.8% to the variance accounted

for, meaning the total LSWH score accounted for a minimal amount of variance after the teacher-awarded grade had been included in the model. It is interesting to note the regression coefficient for the LSWH was negative in the final model. This means that when controlling for teacher-awarded grades, higher LSWH grades predicted *lower* scores on the standardized mathematics examination. Because of the negative regression coefficient, we followed the advice of Smith and colleagues (1992) and tested for suppression effects using semipartial correlations but found no evidence of such effects.

Table 3

Linear Model of Predictors of Standardized Mathematics Examination Scores

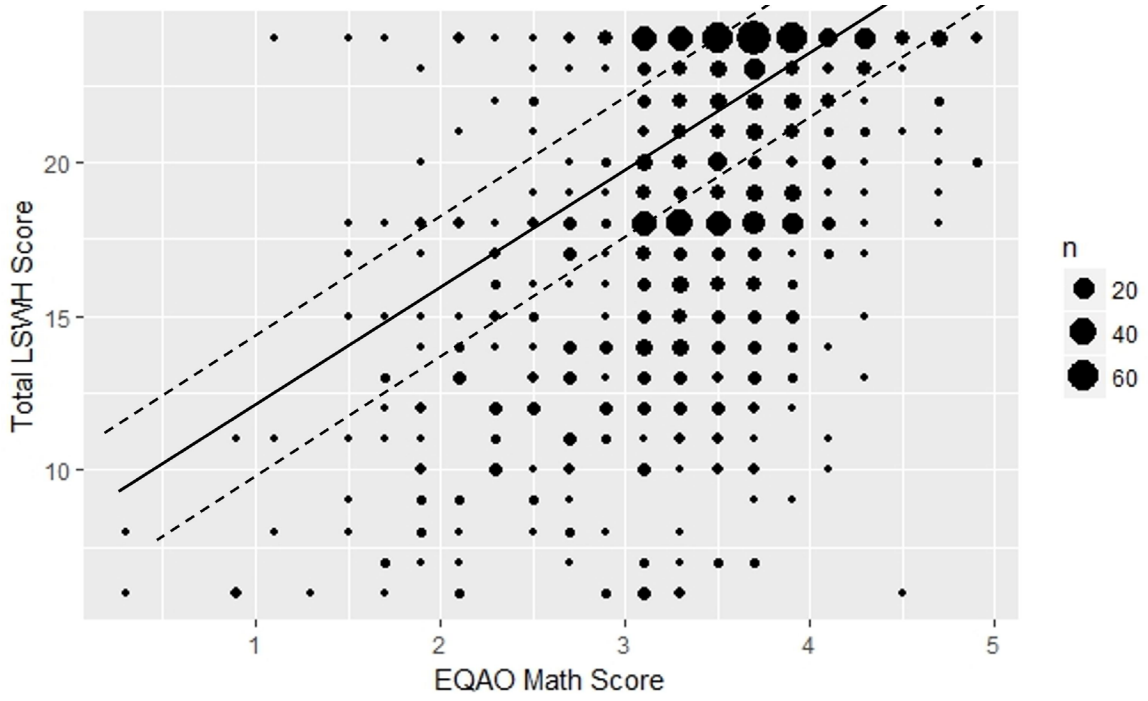
	B	SE B	b	*p*
Step 1				
Constant	3.347	.078		< .001
Gender	.005	.005	.041	.236
Lates	-.042	.014	-.108	.003*
Absences	-.014	.003	-.194	< .001*
Step 2				
Constant	.878	.088		< .001
Gender	-.006	.003	-.043	.055
Lates	-.001	.009	-.002	.936
Absences	-.001	.002	-.007	.766
Teacher-awarded grade	.035	.001	.790	< .001*
Step 3				
Constant	.886	.087		< .001
Gender	-.002	.003	-.019	.412
Lates	-.009	.009	-.022	.339
Absences	-.002	.002	-.023	.319
Teacher-awarded grade	.039	.001	.890	<.001*
LSWH	-.019	.005	-.148	<.001*

Notes:
* Regression coefficient is statistically significant ($p < .05$).
Adjusted $R^2 = .059$ for Step 1; $DR^2 = 0.560$ for Step 2; $DR^2 = 0.008$ for Step 3.
B is the unstandardized regression coefficient, β is the standardized regression coefficient, and p = probability the regression coefficient is due to random chance.
Gender is coded as males = 1, females = 2.

At first glance, the negative regression coefficient for LSWH and standardized mathematics examination scores appears surprising. However, this negative value is small, and is only evident after controlling for the teacher-awarded grade. One possible explanation of the negative effect is that it represents attempts by teachers to compensate some students' poor grades with an acknowledgement of positive effort. We tested this explanation by examining a scatter plot of the LSWH grades vs. the standardized Grade 9 mathematics examination (EQAO) scores (Figure 2). If this explanation were correct, there should be a preponderance of data points with high LSWH grades and low EQAO scores.

Figure 2

Scatterplot of Total LSWH Scores vs. EQAO Score for Grade 9 Mathematics Students in District 2



Note. The dashed lines represent the $p = 0.99$ confidence interval for the line of best fit.

An examination of Figure 2 shows that the data points with the largest number of people have LSWH total scores of 24 and EQAO scores between 3 and 4. These data points fall outside the $p = 0.99$ confidence interval lines for the line of best fit, meaning

these students scored lower on the EQAO mathematics examination than predicted by the line of best fit. Additionally, the figure shows many students who received low LSWH grades did better on the EQAO mathematics examination than predicted.

Girls significantly outperformed boys on all six LSWH. The difference was lower for Grade 12 students than Grade 9 students, but still significant (see Table 4). At both the Grade 9 and 12 levels, girls received higher academic grades than boys across all courses. There was no gender difference in score on the Grade 9 standardized mathematics examination, even though girls received higher teacher-awarded grades in the course than boys ($p < 0.001$). *T*-tests were used to find statistically significant differences between genders, and for all LSWH, girls received higher scores than boys. Table 4 shows the data for District 1 only. The analysis was repeated for District 2 and the numerical results were the same. They are not presented here for reasons of brevity.

Grade 9 physical education is the only course in which there exist separate gender classes. To examine if teachers are norm-referencing their LSWH grades with respect to their classroom norms, the LSWH results for Grade 9 physical education were analyzed separately (see Table 5). If the gender gap disappeared, it would provide evidence that LSWH grades are norm referenced. For this course, the gap between girls' and boys' mean LSWH grade narrowed but did not disappear. Initiative was the only LSWH without a statistically significant difference.

Table 4

Gender Differences in LSWH and Academic Achievement – District 1 Only

	Grade 9				Grade 12			
	Boys (N = 16,131)	Girls (N = 14,182)	p-value	Effect Size	Boys (N = 11,162)	Girls (N = 11,296)	p-value	Effect Size
Collaboration	2.92 (0.01)*	3.26 (0.01)	< 0.001	0.36	3.02 (0.01)	3.29 (0.01)	< 0.001	0.28
Independent Work	2.76 (0.01)	3.19 (0.01)	< 0.001	0.42	2.84 (0.01)	3.20 (0.01)	< 0.001	0.36
Initiative	2.69 (0.01)	3.10 (0.01)	< 0.001	0.39	2.72 (0.01)	3.09 (0.01)	< 0.001	0.35
Organization	2.68 (0.01)	3.22 (0.01)	< 0.001	0.52	2.73 (0.01)	3.18 (0.01)	< 0.001	0.43
Responsibility	2.71 (0.01)	3.15 (0.01)	< 0.001	0.42	2.72 (0.01)	3.11 (0.01)	< 0.001	0.36
Self-Regulation	2.71 (0.01)	3.17 (0.01)	< 0.001	0.45	2.77 (0.01)	3.14 (0.01)	< 0.001	0.35
Teacher-Awarded Grade – All courses**	69.42 (0.13)	75.1 (0.14)	< 0.001	0.34	68.98 (0.18)	74.85 (0.17)	< 0.001	0.31
Teacher-Awarded Grade – Mathematics Only	66.27 (0.39)	69.38 (0.42)	< 0.001	0.18	N/A	N/A	N/A	N/A
Standardized Mathematics Examination ***	3.06 (0.02)	3.03 (0.02)	= 0.34	0.03	N/A	N/A	N/A	N/A

* Number in parentheses is the standard error of the mean.
** Teacher-awarded grades are given as a percentage in Ontario.
*** The standardized mathematics examination is scored on a scale from 0.0 to 4.9. N = 3489.

Table 5

Gender Differences in LSWH and Academic Achievement for Grade 9 Physical Education – District 1 Only

	Grade 9 Physical Education			
	Boys (N = 1,619)	Girls (N = 1,477)	p-value	Effect Size
Collaboration	3.29 (0.02)*	3.40 (0.02)	< 0.001	0.14
Independent Work	3.11 (0.02)	3.26 (0.02)	< 0.001	0.17
Initiative	3.05 (0.02)	3.09 (0.02)	= 0.16	0.05
Organization	2.99 (0.01)	3.20 (0.02)	< 0.001	0.23
Responsibility	3.08 (0.02)	3.21 (0.01)	< 0.001	0.15
Self-Regulation	3.10 (0.02)	3.31 (0.02)	< 0.001	0.24
Teacher-Awarded Grade	76.15 (0.33)	77.87 (0.38)	< 0.001	0.12

* Number in parentheses is the standard error of the mean.

(p = 0.16) in scores between genders. All other LSWH show significant differences (p < 0.001) in mean grade between girls and boys, but effect sizes were smaller than when the LSWH grades were compared from all courses.

Discussion

The analyses of our data highlight several potential issues with teachers' assessment of LSWH. The factor analysis results show the six LSWH grades represent a unidimensional construct, suggesting that teachers are not assessing the six LSWH as distinct constructs. It is possible that the six LSWH are assessed independently of each other, but the constructs themselves are so closely correlated, that the results appear unidimensional. A second pos-

sibility is that teachers are assessing some, or all, of the six LSWH as distinct constructs, but that teachers' definitions of each construct are randomly varied, such that the entire set of grades appears unidimensional. However, given that approximately 80% of the variance in LSWH grades can be accounted for by one factor, and based on our earlier qualitative work, we believe the most likely explanation is that teachers' assessments of the six LSWH are founded upon holistic impressions of their students formed over time. Interviews with Ontario teachers revealed that over half the participants could not name all six LSWH (Merchant, 2016). If teachers are not aware of what the six LSWH are, they cannot be assessing them as separate constructs. Holistic impressions would also explain why approximately half of students receive the same rating across all six LSWH.

We justified testing whether collaboration emerged as a distinct factor because it is the only LSWH that is interpersonal, rather than intrapersonal, and because in prior research teachers reported they were able to separate collaboration as a separate construct, while they were more likely to group the other five LSWH into a single construct. At first glance, the factor analyses provide weak support for this in the report card data, as collaboration always had the lowest loading within the single factor model, regardless of district, grade, course, or stream. However, collaboration also had the highest mean score and lowest standard deviation of the LSWH, and the smaller variance likely reduced factor loadings. When we forced a two-factor model onto the data, the ability to predict the LSWH grades (i.e., how much of the variance in the LSWH grades could be predicted by the model) increased by a minimal amount, and collaboration did not emerge as a separate factor. Thus, it appears that either collaboration is very highly correlated with the other five LSWH, or teachers are not grading it as a separate construct.

Based on the assumption that strong learning skills lead to better learning, it was expected that LSWH grades would correlate positively with academic achievement and standardized test scores. A stronger correlation between LSWH and teacher-awarded grades than with standardized test scores was also expected, based upon prior research demonstrating teacher-awarded grades are not pure measures of achievement, but include subjective judgements, and other factors such as effort and participation (Allal, 2013; Hunter et al., 2006; McMillan, 2001). While it is likely that strong LSWH are positively impacting students' achievement grades, the strength of this association may be influenced by other factors. For example, the direction of influence between achievement and LSWH grades may be bidirectional. It is possible that teachers use achievement grades to inform

their LSWH grades and vice-versa. Another possibility is that because achievement grades and LSWH grades are both assigned by teachers, common method variance is creating strong correlations between the LSWH grades and achievement grades.

The negative regression coefficient for LSWH and standardized mathematics examination scores is surprising. Figure 2 shows many high LSWH students scoring lower on the EQAO examination than predicted, and low LSWH students scoring higher than predicted. This is likely what is responsible for the negative association between LSWH and EQAO scores found in the multiple linear regression. However, it should be re-emphasized that while statistically significant, this coefficient is small. Therefore, the conclusion holds that LSWH scores do not serve as useful predictors of EQAO math achievement above and beyond the teacher-awarded grades. Two possible explanations of the negative coefficient are that teachers compensate some students' poor grades with an acknowledgement of positive effort, or that students with poor LSWH receive lower achievement grades from their teachers.

The results here show that secondary teachers in Ontario rated girls as having better LSWH than boys in all categories. This is true for both Grade 9 and 12, but the difference is smaller in Grade 12. Our results are consistent with Duckworth and Seligman (2006), who found that better self-discipline accounted for girls' achieving higher grades than boys in school. While both our study and that of Duckworth and Seligman measured students' learning skills using teacher judgement, their use of standardized questionnaires avoided the potential of common method variance accounting for strong correlations between self-control measures and achievement grades. In the present study, both the achievement grades and the LSWH grades were determined solely by the teacher, meaning that strong correlations between them may have been partially due to common method variance. One potential explanation for the smaller gender gap in LSWH grades for Grade 12 students is that low achieving males are more likely to drop out than other students (Stetser & Stillwell, 2014; Wang & Fredricks, 2014). It may also be a result of students in Grade 12 becoming more serious about their educational success because of a desire to maximize options for post-secondary education and employment.

Gender gaps in LSWH performance narrow, but do not disappear, in Grade 9 physical education (the only class that is single gender). This suggests that teachers are using a mix of norm- and criterion-referenced frameworks when grading LSWH. Alternatively, it is possible that Grade 9 students demonstrate different behaviours in

single gender classrooms, or that boys have a higher interest in physical education than girls, and this is reflected in improved LSWH grades. While these other explanations are plausible, we contend that in the presence of other research that demonstrates teachers often interpret grading criteria in relation to the norms of their classroom (e.g., Sadler, 2009; Wyatt-Smith & Klenowski, 2013), it is reasonable to conclude that LSWH grades are likely norm-referenced.

Our findings highlight a broader issue with LSWH assessment policies in Ontario. While teachers are given strong guidance and support about how students' academic abilities should progress as they move through the school system, there appears to be minimal guidance or direction about how students' LSWH should develop over time. Ontario provincial curriculum documents and assessment policies do not provide teachers with criteria or standards for the LSWH. Thus, teachers are creating their own internal criteria and standards for the different levels of LSWH performance at each grade level. We do not know, for example, how teachers distinguish between "excellent" collaboration and "good" collaboration, nor whether those standards are different for Grade 12 students than for Grade 9 students. More research needs to be done to understand how teachers define these constructs, what activities they use to assess them, and what student behaviours and characteristics influence the grades awarded.

Implications and Limitations

Stecher and Hamilton (2014) found that while teachers are interested in developing and assessing learning skills, they "do not have the resources to develop programs or assessments on their own" (p. 7). Our findings here provide further empirical evidence of the necessity to support teachers' classroom assessment practices in relation to learning skills. Given their sustained interaction with students over time, and in a variety of contexts, teachers are well positioned to be good assessors and reporters of students' learning skills (Zimmerman & Martinez-Pons, 1988). However, our evidence strongly suggests that teachers' grades of such skills are likely based more upon holistic impressions of the student than on performance standards and well-defined constructs. Consistent with this hypothesis is our finding that approximately half of students receive the same grade across all six LSWH. Such findings are not surprising given that very few teachers receive training or guidance on how to define, assess, and grade such skills. In Ontario, there is no requi-

rement for pre-service teacher education to include courses or even a single lesson on how to assess the LSWH. This points to the need for teachers to be given training, and concrete examples of how to assess learning skills in a manner that is defensible and valid. School systems interested in having teachers assess and report upon learning skills need to give careful consideration as to what skills should be assessed, how they can be defined clearly for teachers, students, and parents, and appropriate standards for these skills at different grade levels. Teachers will need to have resources in place including ongoing training, sample assessment activities, and grade level standards.

Even with such training, it is not clear that teachers would assess the LSWH as independent constructs. Evidence from the rater training literature demonstrates inconsistent effects on how effective rater training is at reducing halo effects (e.g., Bernardin, 1978; Woehr & Huffcutt, 1994), and that the effects of rater training diminish over time (Ivancevich, 1979). One possible way to prevent halo effects from impacting teachers' LSWH grades would be to stop asking teachers to assess multiple LSWH for the same student. It would be possible in secondary schools for different subjects to be responsible for assessing and reporting upon a single LSWH. Because students generally take more than six subjects during a school year, all six LSWH could be assessed and reported for most students.

Girls obtaining higher grades than boys is a well-known phenomenon that is documented across grade levels, races, and contexts (Voyer & Voyer, 2014). Our research finds that, in Ontario, this gender gap extends to non-achievement constructs such as the LSWH. This finding is consistent with the findings of Duckworth and Seligman (2006), although the results shown here demonstrate a smaller gender gap than they found. Our results add to the growing body of research indicating that boys receive lower grades than girls, but adds evidence that boys also receive lower grades for non-achievement constructs such as the Ontario LSWH. Our finding that gender differences in standardized mathematics scores are minimal is consistent with other data such as the Canadian PISA results (Brochu et al., 2013). The lack of significant gender differences in standardized mathematics examination scores suggests that boys and girls have equal mathematical ability, and so girls' higher achievement on report cards is due to other factors. One possibility is that teachers perceive girls to have better LSWH, and this leads to better grades as teachers consider the LSWH when assigning achievement grades. This explanation is consistent with prior research on teachers' grading practices that

demonstrates teachers include non-achievement factors in their achievement grades (e.g., Bowers, 2011; McMillan, 2001). Another possibility is that teachers are biased against boys in their grading decisions. Protivínský and Münich (2018) reviewed 13 studies related to gender bias in teacher grading and reported that 11 of the 13 found a grading bias against boys. Falch and Naper (2013) suggest that gender differences in student-teacher interactions are responsible for gender differences in achievement grades. If so, it seems likely that differences in student-teacher interactions would also lead to gender differences in LSWH grades.

The most severe limitations of this study are imposed by the nature of the data obtained. While the large number of data points improves statistical power, the limited amount of information collected for each set of grades constrained the analyses that could be performed. As an example, it would be interesting to test how variable the LSWH grades are across contexts. Are the biggest sources of variance the student, the teacher, or the school? Unfortunately, the anonymity of the data prevents us from answering these questions. It would also be interesting to look at standardized test scores for subjects other than mathematics, as mathematics appears to be the subject where non-achievement factors have the smallest impact on achievement grades (Bol et al., 1998; Duncan & Noonan, 2007; Pilcher, 1994). Further, it would be helpful to complement the data with different ratings of the LSWH or related constructs such as conscientiousness or self-control. These additional measures could help determine the extent to which strong correlations between achievement and the LSWH grades are due to common method variance.

Conclusion

Ontario secondary teachers assess the six LSWH as a unitary construct, which accounts for over 80% of the variance in the LSWH. This indicates that LSWH grades do not reflect separate performance levels on six distinct constructs, but a teacher's overall impression of the student. Consistent with other research, we found LSWH grades are more strongly correlated to teacher-awarded grades than standardized test scores. This implies that teachers conflate achievement and non-achievement factors when grading, although other explanations, such as common method variance, exist. Gender gaps in LSWH achievement are significant, both statistically and practically, and it would be worthwhile to examine further the impact this has on gender gaps in academic achieve-

ment, and whether improving boys' LSWH would help close this gap. Taking a broader perspective, these findings from Ontario imply that if school systems wish to incorporate the assessment and grading of learning skills at the classroom level, they need to provide teachers with appropriate supports, including training, sample assessment activities, and grade level standards.

References

Allal, L. (2013). Teachers' professional judgement in assessment: A cognitive act and a socially situated practice. *Assessment in Education: Principles, Policy & Practice*, *20*(1), 20–34.

Almlund, M., Duckworth, A. L., Heckman, J. J., & Kautz, T. D. (2011). *Personality psychology and economics* (No. w16822). National Bureau of Economic Research.

Association for Supervision and Curriculum Development. (2007). *The learning compact redefined: A call to action*. http://www.ascd.org/ASCD/pdf/Whole%20Child/WCC%20Learning%20Compact.pdf.

Babad, E. Y., Inbar, J., & Rosenthal, R. (1982). Teachers' judgment of students' potential as a function of teachers' susceptibility to biasing information. *Journal of Personality and Social Psychology*, *42*(3), 541–547.

Bass, K. M. (2005). Reality's limits. *Measurement: Interdisciplinary Research and Perspectives*, *3*(2), 84–88.

Baumeister, R. F., & Vohs, K. D. (2007). Self-regulation, ego depletion, and motivation. *Social and Personality Psychology Compass*, *1*(1), 115–128.

Bernardin, H. J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. *Journal of Applied Psychology*, *63*(3), 301–308.

Bol, L., Stephenson, P. L., O'connell, A. A., & Nunnery, J. A. (1998). Influence of experience, grade level, and subject area on teachers' assessment practices. *The Journal of Educational Research*, *91*(6), 323–330.

Borghans, L., Duckworth, A. L., Heckman, J. J., & Ter Weel, B. (2008). The economics and psychology of personality traits. *Journal of Human Resources*, *43*(4), 972–1059.

Bowers, A. J. (2011). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. *Educational Research and Evaluation*, *17*(3), 141–159.

Briesch, A. M., Chafouleas, S. M., & Riley-Tillman, T. C. (2010). Generalizability and dependability of behavior assessment methods to estimate academic engagement: A comparison of systematic direct observation and direct behavior rating. *School Psychology Review*, *39*(3), 408–421.

Brochu, P., Deussing, M. A., Houme, K., & Chuy, M. (2013). *Measuring up: Canadian results of the OECD PISA study*. Council of Ministers of Education.

Brookhart, S. M. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement*, *30*(2), 123–142.

Brown, G. T. (2011). Teachers' conceptions of assessment: Comparing primary and secondary teachers in New Zealand. *Assessment Matters*, *3*, 45–70. https://link.gale.com/apps/doc/A307413786/AONE?u=queensulaw&sid=bookmark-AONE&xid=f122ed65

Casner-Lotto, J. & Barrington, L. (2006). *Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century US workforce*. Partnership for 21st Century Skills.

Cizek, G. J. (1996). Grades: The final frontier in assessment reform. *NASSP Bulletin, 80*(584), 103–10.

Conference Board of Canada. (2015). Employability Skills 2000+. http://www.conferenceboard.ca/topics/education/learning-tools/employability-skills.aspx

Cross, L. H., & Frary, R. B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Education*, *12*(1), 53–72.

DeLuca, C., & Bellara, A. (2013). The current state of assessment education aligning policy, standards, and teacher education curriculum. *Journal of Teacher Education*, *64*(4), 356–372.

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168.

Dinsmore, D. L., Alexander, P. A., & Loughlin, S. M. (2008). Focusing the conceptual lens on metacognition, self-regulation, and self-regulated learning. *Educational Psychology Review*, *20*, 391–409.

Duckworth, A., & Gross, J. J. (2014). Self-control and grit: Related but separable determinants of success. *Current Directions in Psychological Science*, *23*(5), 319–325.

Duckworth, A., & Seligman, M. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades and achievement test scores. *Journal of Educational Psychology*, *98*(1), 198–208.

Duckworth, A. L., & Schulze, R. (2009). *Jingle jangle: A meta-analysis of convergent validity evidence for self-control measures* [Manuscript]. University of Pennsylvania, Department of Psychology.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*(4), 237–251.

Duncan, C. R., & Noonan, B. (2007). Factors affecting teachers' grading and assessment practices. *Alberta Journal of Educational Research*, *53*(1), 1–21.

Falch, T., & Naper, L. R. (2013). Educational evaluation schemes and gender gaps in student achievement. *Economics of Education Review*, *36*, 12–25. https://doi.org/10.1016/j.econedurev.2013.05.002

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.* University of Chicago Consortium on Chicago School Research.

Ferrito, J. (2015). *An investigation of the relationship between behavioral feedback on social-emotional skills on report cards and academic achievement* [Unpublished doctoral dissertation]. Rutgers, The State University of New Jersey.

Fieldwork Education (n.d.). *International Primary Curriculum*. https://fieldworkeducation.com/curriculums/primary-years

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906–911.

Galla, B. M., Plummer, B. D., White, R. E., Meketon, D., D'Mello, S. K., & Duckworth, A. L. (2014). The Academic Diligence Task (ADT): Assessing individual differences in effort on tedious but important schoolwork. *Contemporary Educational Psychology*, *39*(4), 314–325. https://doi.org/10.1016/j.cedpsych.2014.08.001

Greene, J. A., Robertson, J., & Costa, L. J. C. (2011). Assessing self-regulated learning using think-aloud methods. In B. Zimmerman & D. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 313–328). Routledge.

Guskey, T. R., Swan, G. M., & Jung, L. A. (2011). Grades that mean something: Kentucky develops standards-based report cards. *Phi Delta Kappan*, *93*(2), 52–57.

Hadwin, A., & Winne, P. (2012). Promoting learning skills in undergraduate students. In J. R. Kirby & M. J. Lawson (Eds.), *Enhancing the quality of learning: Dispositions, instruction, and learning processes* (pp. 315–338). Cambridge University Press.

Howley, A., Kusimo, P. S., & Parrott, L. (2000). Grading and the ethos of effort. *Learning Environments Research*, *3*(3), 229–246.

Hunter, D., Mayenga, C., & Gambell, T. (2006). Classroom assessment tools and uses: Canadian English teachers' practices for writing. *Assessing Writing*, *11*(1), 42–65.

International Baccalaureate Organization. (2009). *IB learner profile booklet*. http://www.ibo.org/programmes/documents/learner_profile_en.pdf

Ivancevich, J. M. (1979). Longitudinal study of the effects of rater training on psychometric error in ratings. *Journal of Applied Psychology*, *64*(5), 502–508.

Ivcevic, Z., & Brackett, M. (2014). Predicting school success: Comparing conscientiousness, grit, and emotion regulation ability. *Journal of Research in Personality*, *52*, 29–36.

Jacob, B. A. (2002). Where the boys aren't: Non-cognitive skills, returns to school and the gender gap in higher education. *Economics of Education Review*, *21*(6), 589–598.

Keeley, J. W., English, T., Irons, J., & Henslee, A. M. (2013). Investigating halo and ceiling effects in student evaluations of instruction. *Educational and Psychological Measurement*, *73*(3), 440–457.

Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural equation modeling*, *10*(3), 333–351.

Kleitman, S., & Costa, D. S. (2014). The role of a novel formative assessment tool (StatsmIQ) and individual differences in real-life academic performance. *Learning and Individual Differences*, *29*, 150–161.

Lai, E. R. (2011). *Metacognition: A literature review*. images.pearsonassessments.com/images/tmrs/Metacognition_Literature_Review_Final.Pdf

Lekholm, A. K., & Cliffordson, C. (2008). Discrepancies between school grades and test scores at individual and school level: Effects of gender and family background. *Educational Research and Evaluation*, *14*(2), 181–199. https://gupea.ub.gu.se/bitstream/handle/2077/18673/gupea_2077_18673_2.pdf;jsessionid=BD7AB2B4C5A13D83177AD4FD95AADA66?sequence=2

Levin, H. (2012). More than just test scores. *Prospects*, *42*(3), 269–284.

Linn, R., & Miller, M. (2005). *Measurement and assessment in teaching*. Pearson Prentice Hall.

McGill, D. A., Van der Vleuten, C. P. M., & Clarke, M. J. (2011). Supervisor assessment of clinical and professional competence of medical trainees: A reliability study using workplace data and a focused analytical literature review. *Advances in Health Sciences Education*, *16*, 405–425.

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice*, *20*(1), 20–32.

McMillan, J. H., & Workman, D. J. (1998). *Classroom assessment and grading practices: A review of the literature.* Metropolitan Educational Research (ERIC Document Reproduction Service No. ED453-63).

Merchant, S. (2016, September 29–October 1). *Assessment and reporting of learning skills and work habits: Capturing the complexity* [Conference presentation]. The Annual Conference of the Consortium for Research on Educational Assessment and Teaching Effectiveness, Louisville, KY, United States.

Merchant, S., Klinger, D. A., & Love, A. (2018). Assessment and reporting of non-cognitive skills: A cross-Canada survey. *Canadian Journal of Educational Administration and Policy*, *187*, 2–17.

Miller, T., Klinger, D., & Shulha, L. (2006). Behaviour assessment in Ontario mathematics classrooms. *Educational Research and Reviews*, *1*(1), 1–6.

Muenks, K., Wigfield, A., Yang, J. S., & O'Neal, C. R. (2017). How true is grit? Assessing its relations to high school and college students' personality characteristics, self-regulation, engagement, and achievement. *Journal of Educational Psychology*, *109*(5), 599–620.

National Education Association. (2012). *Preparing 21st century students for a global society: An educator's guide to the "four Cs."* National Education Association.

Northern Ireland Ministry of Education. (2007). Thinking skills and personal capabilities for key stage 3. Council for the Curriculum, Examinations & Assessment. https://ccea.org.uk/key-stage-3/curriculum/thinking-skills-personal-capabilities

Ontario Ministry of Education. (2005). *The Ontario curriculum grade 9 and 10 mathematics*. http://www.edu.gov.on.ca/eng/curriculum/secondary/math910curr.pdf

Ontario Ministry of Education. (2010). *Growing success: Assessment, evaluation and reporting in Ontario schools*. http://www.edu.gov.on.ca/eng/policyfunding/growsuccess.pdf

Ontario Ministry of Education. (2016). *21st century competencies: Foundation document for discussion*. http://www.edugains.ca/resources21CL/21stCenturyLearning/21CL_21stCenturyCompetencies.pdf

Pellegrino, J., & Hilton, M. (2013). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. National Academies Press.

Pilcher, J. K. (1994). The value-driven meaning of grades. *Educational Assessment*, *2*(1), 69–88.

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*(1), 33–40.

Poth, C. A. (2012). What assessment knowledge and skills do initial teacher education programs address? A Western Canadian perspective. *Alberta Journal of Educational Research*, *58*(4), 634–656.

Protivínský, T., & Münich, D. (2018). Gender bias in teachers' grading: What is in the grade? *Studies in Educational Evaluation*, *59*, 141–149.

Russell, J. A., & Austin, J. R. (2010). Assessment practices of secondary music teachers. *Journal of Research in Music Education*, *58*(1), 37–54.

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, *34*(2), 159–179.

Schunk, D. H. (2008). Metacognition, self-regulation, and self-regulated learning: Research recommendations. *Educational Psychology Review*, *20*, 463–467.

Silva, E. (2009). Measuring skills for 21st-century learning. *Phi Delta Kappan*, *90*(9), 630–634.

Singapore Ministry of Education. (2021). *Character and citizenship education syllabus (CCE) – Secondary*. https://www.moe.gov.sg/-/media/files/secondary/syllabuses/cce/2021-character-and-citizenship-education-syllabus-secondary.pdf

Smith, R., Ager, J., & Williams, D. (1992). Suppressor variables in multiple Regression/Correlation. *Educational and Psychological Measurement, 52*(1), 17–29.

Sparks, S. (2014). New character report cards rate students on 'grit.' *Education Week*. http://www.edweek.org/ew/articles/2014/06/05/34measuring-motivation-b4.h33.html

Stecher, B. M., & Hamilton, L. S. (2014). *Measuring hard-to-measure student competencies: A research and development plan*. RAND.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245–251.

Stetser, M. C., & Stillwell, R. (2014). *Public high school four-year on-time graduation rates and event dropout rates: School years 2010-11 and 2011-12. First Look. NCES 2014-391*. National Center for Education Statistics.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*(1), 25–29.

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, *140*(4), 1174–1204.

Wang, M. T., & Fredricks, J. A. (2014). The reciprocal links between school engagement, youth problem behaviors, and school dropout during adolescence. *Child Development*, *85*(2), 722–737.

Watson, D. L., & Emery, C. (2010). From rhetoric to reality: The problematic nature and assessment of children and young people's social and emotional learning. *British Educational Research Journal*, *36*(5), 767–786.

Wigelsworth, M., Humphrey, N., Lendrum, A. & Kalambouka, A. (2010). A review of key issues in the measurement of children's social and emotional skills. *Educational Psychology in Practice*, *26*(2), 173–186.

Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeider (Eds.), *Handbook of self-regulation* (pp. 531–566). Academic Press.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, *67*, 189–205.

Wyatt-Smith, C., & Klenowski, V. (2013). Explicit, latent and meta-criteria: Types of criteria at play in professional judgement practice. *Assessment in Education: Principles, Policy & Practice*, *20*(1), 35–52.

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist*, *25*(1), 3–17.

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeider (Eds.), *Handbook of self-regulation* (pp. 13–39). Academic Press.

Zimmerman, B. J. (2013). From cognitive modeling to self-regulation: A social cognitive career path. *Educational Psychologist*, *48*(3), 135–147.

Zimmerman, B. J., & Kitsantas, A. (2014). Comparing students' self-discipline and self-regulation measures and their prediction of academic achievement. *Contemporary Educational Psychology*, *39*(2), 145–155.

Zimmerman, B. J., & Martinez-Pons, M. (1988). Construct validation of a strategy model of student self-regulated learning. *Journal of Educational Psychology*, *80*(3), 284–290.