visible learning plus

Visible Learning International Impact Report

› Technical supplement

June 2015

This report was written by:

Dr Heidi Leeson - External Education Evaluator
Monocle Solutions Ltd
heidi.leeson@monocle.co.nz

Overview

The rationale for the creation of the *Visible Learning*^{plus} *International Impact Report* was to use empirical evidence to measure empirically, the impact of the Visible Learning^{plus} program. The various methodological approaches that were used to analyze this impact are explained in this document. Collectively, these approaches enabled robust approaches to be used in assessing the impact of the Visible Learning^{plus} program across various education systems, schools and students. While a variety of international data sets were utilized, the techniques allowed meaningful findings to be attributable to the impact of the Visible Learning^{plus} program.

Evidence relating specifically to the Visible Learning^{plus} workshops included all the workshops that have been delivered across eight countries. Data relating to the intermediate, 3-9 months, and yearly outcomes were compiled specifically from Australia and New Zealand, where existing and developing evaluation tools have formed part of the system-wide projects (e.g., Collaborative Impact Visible Learning^{plus} Program). From these various datasets a mixture of cross-sectional, repeated measures, longitudinal the approaches were used to measure impact aross the components represented in the evaluation logic model. Where additional analysis was conducted in preparation of the findings presented in the report, e.g., checking for differences in student achievement based on demograhics, were conducted, these are included in each components data analysis section.

The logic model presented in the *Visible Learning Impact Report* (2015), showed schematically the links between the various linear components of the program and the indictors of the impact for each of these as schools progress across a 12-month implementation period. For clarification, the *Visible Learning*^{plus} logical model is based on Kirkpatricks evaluation model (1998). Kirkpatrick's model provides a useful framework for presenting a linear progression of the programs implementation, and related empirical indicators. This framework supports the implementation of the program model which of course, is based on Hattie's research and meta-analyses focused on "what works best" in education systems and from this, what impacts most on students learning and achievement at school.

Significance of the Research

Although a vast array of empirical evidence of the impact and of the Visible Learning^{plus} program have been conducted over the last five years, the *International Impact Report* (2015) provided the first aggregation of data from across various programs. Analysis at this large scale across different education settings allows more generalizability assertions to be made about the impact of the program within different populations. The findings establishes the first step in the creation of a series of reports produced annually, that will confirm (or refute) the interpretations made from this first report. Further, increased datasets will provide the

opportunity to increase the types of analyses that can be conducted on the data. For example, it is proposed that towards the end of 2015, two large Collaborative Impact Visible Learning^{plus} Program [1] will have sufficient data to permit a hierarchical linear model to be created expressly to test the existence of a multilevel evaluation model. This would account for student achievement against school contextual (School Capability Assessments), and teacher attitudes and characteristics (Mindframes Survey) and practice variables (Classroom Observation Tool). As these two programs have also introduced system-level evaluation (i.e., Ministry of Education) a potential 4-level causal models can be tested. This would provide significant data regarding the effects occurring at each level and across levels occurring within education systems implementing the Visible Learning^{plus} and Collaborative Impact Programs. Additional variables that are also collected, such as the geo-location (e.g., urban or rural) or socio-economic/poverty indices can also be in brought into the model design (see Adcock & Phillips, 1997).

The following sections present the methods that were used to analyze the empirical data collected for the measures used as part of the Visible Learning^{plus} program. Each of these measures will be presented separately within the methods section, in the order that they appear in the *International Impact Report* (2015).

[1] The Collaborative Impact Program (CIP) combines the principals of Visible Learning with the Culture Counts relationship-based paradigm. For more information on the CIP, please refer to www.visiblelearning.com.

Methods – Visible Learning^{plus} Workshops

Participants

Approximately 7,500 workshop participants from 13 countries (Australia, Sweden, Denmark, Norway, Belgium, Japan, the Netherlands, New Zealand, Thailand, Indonesia, Malaysia, the United Kingdom, and the United States) completed evaluations. Participants were a mixture of teachers, school leaders and Impact Coaches depending on whether the workshop is applicable to both roles or is role specific.

The evaluations were collected from various Visible Learning^{plus} Foundation Series workshops including Foundation Day, Visible Learning into Action for Teachers 1 & 2, Evidence into Action 1 & 2, and the Inside Series which consisted of Using Data to Know Your Impact, Feedback to Make Learning Visible, Creating Effective Assessments for Teaching and Learning Using The SOLO Taxonomy, and Building and Developing Visible Learners.

Measure

The workshop evaluation assesses four key indicators concerning the quality of the output, specifically, content usefulness, material quality, motivation, and facilitator delivery. The workshop evaluations questionnaires range in length from 10- to 20-items, depending on the workshop. The information from the evaluations focusses explicitly on testing the level of participants understanding of the Visible Learning program theory. Items on all areas are rated on a 6-point Likert scale (1 = very *strongly disagree*; 2 = *strongly disagree*; 3 = *disagree*; 4 = *somewhat agree*; 5 = *agree*; 6 = *strongly agree*), with the 'agreed' response categories aggregated. The Visible Learning^{plus} program sets the benchmark standard of 80% applied across these three aggregated categories.

Procedure

Workshop evaluations are distributed at the end of the workshops. Typically, the results of this feedback are presented in a report that is sent back to the participant's schools.

Data Analysis

Quantitative information was analyzed using proportional information and significance testing of gains made before and after workshops. There were no statistically significant differences between the participants' role in the school and their workshop responses. Similarly, there were no differences found in degree of gains in understanding that occurred before and after the respective workshops. Text and document analysis has been used to establish the patterns and trends in responses from the open-ended sections of the workshop evaluations.

Methods – School Capability Assessment

Participants

There were two groups of participants for the analysis conducted on the school capacility assessment data. Cross-sectional analysis used Australian and New Zealand from assessments conducted in 2014. This sample consisted on approximately 15 schools. The second group of 110 schools, were participants from the Collaborative Impact Program from 2012 – 2014. This sample was suitable for assessing the longitudinal impact of the program.

Measures and Procedure

Schools received two half-day site visits from the Visible Learning^{plus} consultants. The purpose of these visits is to evaluate schools' current practices against the characteristics outlined in the four Visible Learning^{plus} strands (Time 1) and then again at the end of the year (Time 2). Evaluation at these two time periods allows schools and the Visible Learning^{plus} team to measure the impact of the Visible Learning^{plus} program at the school level

A matrix containing focus questions for each Visible Learning^{plus} strand enables the consultants to analyze the schools' systems systematically. In addition, supporting documents facilitate semi-structured interviews with school leaders, teachers, and students, and conduct qualitative observations of a sample of classroom interactions.

This process facilitates the understanding of both the consultant and the school regarding which of their current practices are already meeting some or all of the characteristics within each strand. Against the categories of vision and values, knowledge and understanding, personal qualities, and professional practices, the consultants rate the degree to which evidence of the four strands are present.

Data Analysis

To ensure that it was appropriate to aggregate all the school capability ratings, the assumptions of repeated measures ANOVAs were tested, specifically, independence of observations, normality, and homogeneity of variances (i.e., sphericity). These assumptions with the exception of sphericity were established, with epsilons of less than 1.0 indicating that the assumption of sphericity was violated. Although ANOVAs are reasonably robust to this violation, a correction test (Greenhous-Geisser) was applied when conducting the repeated measures ANOVA. Longitudinal results indicated that there were statistically significant increases in school capabilities over each of the three years of participation in the Collaborative Impact Program across all of the strands. ANOVAs were used to examine the gains made over one year's involvement in the program. Gains analysis was conducted based the changes in consultants' ratings across the two time periods. Findings showed that schools significantly increased their capability to make the Visible Learning^{plus} strands commonplace

and embedded within their schools. Text and document analysis was conducted on both the cross-sectional and longitudinal data sets.

Methods – Mindframes Survey

Participants

The survey was administered to senior leadership and teachers (*n* = 1050) from schools in New Zealand and Australia.

Measure

The Mindframes Survey is an 58-item measure that can be separated into the following 12 distinct constructs: Fixed vs Growth (change agent) (5 items); Focus on learning (7 items); Student voice (3 items); Know thy impact (5 items); High expectations for all (5 items); Dialogue, not monologue (5 items); Trust in class (5 items); Errors are welcome (5 items); Teaching is to DIE for (5 items); Awareness of growth (5 items); Challenge vs do your best (4 items); Assessment for teaching (4 items). Items on all areas are rated on a positively weighted 6-point Likert scale (1 = *strongly disagree*; 2 = *disagree*; 3 = *somewhat agree*; 4 = *agree*; 5 = *strongly agree*; 6 = *very strongly agree*). The positively weighting design was a result of previous analysis which indicated a need for greater discrimination between the *agreed* response options.

Procedure

Over a school year, data is collected at two time periods: initial baseline (Time 1) at the beginning of the year and at the end of the year (Time 2). The Time 1 administration occurs at the commencement of the Foundation Day workshop, prior to any significant exposure of any Visible Learning theory. Thus, responses at this time period will reflect the pedagogical beliefs and approaches currently held by teachers and school leaders. Responses at Time 2 reflect the change that has occurred in participants mindframes through the knowledge, understanding and implementation of Visible Learning^{plus} over the school year.

Data Analysis

Mindframes Survey feedback used for the analysis was conducted on data that had been collected for the two time periods across the last

A total of 11% and 14% of data were missing, respectively for the two data collection periods. Research has demonstrated that missing data methods such as pairwise, listwise, or mean imputations are inappropriate to use, as they tend to yield biased estimates (Little & Rubin, 1987; Wothke, 2000). Little and Rubin (1987) suggested that when analyzing repeated

measures data with missing data, it should consider whether it is missing completely at random (MCAR) or missing at random (MAR). The assumption of MCAR is that the pattern of missing values does not depend on the missing value. In other words, there is nothing systematically related to the data being missing. MCAR is the more restrictive in that it assumes the missing data are completely independent of all the variables in the model. MAR, on the other hand, is less restrictive, in that it assumes that the missing data may be related to the observed data. Following Conroy, Metzler, and Hofer's (2003) example of examining "missingness" of data, it was found the number of missing waves of data not to be correlated ($p < .05$) to total scores on the Mindframes Survey at each time point. This finding showed that the missing data at Time 2 was not deemed to be related to the Mindframes Survey scores at Time 1, given the data were MAR.

Exploratory Factor Analysis of the Mindframes Survey

Establishing the factor structure of the Mindframes Survey is important as it indicates whether the structure of the scale concurs with the expected constructs that it intends to measure. To assess the structure of the Mindframes Survey, a confirmatory factor analysis (EFA) was conducted with the aim of confirming the number of common factors influencing this measure. As the assumption of multivariate normality was satisfied, factor analysis utilized the maximum likelihood method for extraction with an oblique rotation. This rotation method was selected based on the desire to extract factor patterns regardless of their degree of correlation, and based on previous evidence which has suggested that the constructs in the Mindframes are not orthogonal. Therefore, an oblique rotation approach presented a more precise and realistic depiction of how the 12 constructs in the Mindframes Survey are likely to be related to one another (Fabrigar, Wegener, MacCallum & Strahan, 1999).

Two rules, namely, Cattell's (1966) scree plot and Kaiser's (1960) eigenvalues rule (eigenvalues >1) were used to determine the latent structure of the Mindframes Survey. The scree plot of eigenvalues showed a clear break in the data, with eight factors positioned above this break. However, given that 12 factors had eigenvalues larger than 1.00, accounting for 69.24% of the variability; a 12-factor solution was confirmed as being the most parsimonious explanation of the data.

Reliability Analysis

Reliability estimates (Cronbach's alpha) were used to evaluate the stability of the Mindframes Survey across the two administration periods (test-retest reliability) and the equivalence of the items associated with each of the 12 constructs and overall (internal consistency).

The strength of the association of the two sets of scores (Time 1 and Time 2) were strong (r = > .70). Similarly, the alpha coefficients ranged from .71 to 92 for each of the Mindframes constructs, with an overall survey alpha of α = .87 (Time 1) and α = .90 at (Time 2).

Gains Analysis

A mixed ANOVA was conducted to assess whether there were school role (school leader or teacher) and school characteristics (e.g., primary or secondary, region, country) variables impacting on the Mindframes Survey responses. Results indicated that no significant main effects where found for either school role ($p > .05$) or any of the school characteristics ($p > .05$). There were no significant interactions between either independent variables or Mindframes Survey responses. These multivariate results indicated that there were no statistical reasons hindering the aggregation of Mindframes data for repeated measures and gains analyses.

To ensure that it was appropriate to aggregate all Mindframes Survey responses, the assumptions of repeated measures ANOVAs were tested, specifically, independence of observations, normality, and homogeneity of variances (i.e., sphericity). These assumptions with the exception of sphericity were established, with epsilons of less than 1.0 indicating that the assumption of sphericity was violated. Although ANOVAs are reasonably robust to this violation, a correction test (Greenhous-Geisser) was applied when conducting the repeated measures ANOVA. Results indicated that there were statistically significant increases in participants Mindframes responses the one-year period for all of the 12 construct areas. Examination of the means suggested that the greatest gain (+1.12) occurred for the Know thy Impact construct. Polynomial contrast indicated that there was a significant linear trend, F (1, 1050) = 38.69, $p < .001$, $eta^2 = .84$.

Methods – Classroom Observation Tool

Participants

A total of 16 Impact Coaches conducted 46 classroom observations - 25 in urban schools and 21 in remote schools. Nearly half of the observations were conducted in English (including literacy and reading) classes, with the others during mathematics (28%) or science (7%) classes. The majority of observations were carried out in primary and middle schools (80%) with senior schools representing 20% of the classrooms observed (Years 10 - 11). Classroom observations over two or three-time series allow the assessment of the degree to which practices and teacher-student interactions change as the school evolves in their professional learning and implementation of the Visible Learning^{plus} strands.

Measure

The primary purpose of the Classroom Observation Tool is to enhance learning for the teacher and track how the teacher is improving his/her practices to have a greater impact on student learning, engagement, progress and achievement. The Classroom Observation Tool consists of three parts, with each part represents different aspects of the Visible Learning^{plus} program. Part 1 is focused finding evidence of Visible Learners. Visible Learners are defined

as students who actively engage in their learning. Two parts of the Visible Learningplus Classroom Observation Tool focus on capturing evidence of this behaviour occurring amongst students in the classroom environment. Part 1 directly reflects specific characteristics that illustrate a visible learner as identified by the Impact Coach. Whereas in Part 3, Impact Coaches ask students for their own perceptions and understandings of where they are at, how they are doing and where they are going next.

The six characteristics of visible learner behaviour are:

1. Asking the teacher and other students questions
2. Talking about what they are learning
3. Voicing and demonstrating high expectations
4. Supporting their peers learning
5. Seeking feedback from the teacher and other students
6. An awareness of their learning steps

The Impact Coach measures each of these characteristics by responding to both a 4-point Likert scale (1 = *no evidence present*; 2 = *slight evidence*; 3 = *moderate evidence*; 4 = *substantial evidence*), and by recording the behaviour that they have rated (open-ended comments).

The goal of Part 2 of the Classroom Observation Tool was to examine the interactions that a teacher has within their classroom learning context. At the beginning of Part 2, the Impact Coach selects the Visible Learning strand and associated characteristics that the teacher, during the pre-observation meeting, had requested to be the focus of the observation. The once the class commences, the Impact Coach spends the following 15 minutes noting the different instances where the teacher's behaviour and interactions are representing the chosen strand's characteristics. Results showed that typically the Impact Coaches noted any of the Visible Learning characteristics that they saw occurring during this time. Whilst observing the teachers interactions with students, the Impact Coaches also note the location of the teacher in the classroom. Specifically, against each characteristic was the teacher sitting at their desk, stationary, moving around the room, or out of the room?

As mentioned previously above, Visible Learners are defined as students who are actively engaged in their learning. In addition to the Impact Coach capturing evidence of visible learners in Part 1 of the tool, Part 3 is focused on getting evidence of student's engagement their learning. This is achieved by asking students three questions: what are you learning today; how do you know how well you are going; and what do you think your next steps are?

Procedure

As part of the evaluation of the Collaborative Impact Visible Learning^{plus} Program schools are required to provide evidence of teacher practice using the Classroom Observation Tool.

Over a school year, data is collected at two time periods: initial baseline (Time 1) at the beginning of the year and at the end of the year (Time 2). As such, the expectation is for a teacher to undergo a minimum of two observations in the same school year. Clear guidelines and training accompany the tool so that the protocols surrounding its use help standardize the methodology and as a result, the reliability of the results. Pre- and post-observation meetings are conducted between the school leader/Impact Coach and the teacher. At the pre-observation meeting the teacher is introduced to the tool and invited to nominate the Visible Learningplus strands and characteristics her/he is currently focusing her/his learning and teaching practices on. The post-observation meeting allows for the both to discuss the findings from the observation and make judgements on his/her practice and any changes he/she might which to explore, or make, in relation to their practice.

Data Analysis

The following analyses are presented in the three parts of the classroom observation tool. All three represent different aspects of the Visible Learning program, are measured using different techniques, with evidence of impact being observed across both the teachers and students behaviours and interactions.

Parts 1 & 2 – The Visible Learner

MANOVAs were conducted across various demographics such as school type (primary, intermediate or secondary/high school; urban and rural, year levels) against the 6-point Likert scale items. No significant differences were found across any of these dependent variables.

For analysis of the open-ended comments text and sentiment analytics was conducted in order to ascertain the range and differentiation of visible learner behaviors'. Text mining was conducted using *R* programing language specifically developed for analyzing text. In this context, text analytics was applied to extract particular nuances from the language used by the Impact Coaches when describing the students' behaviour, specifically looking for frequency and relationships between different words and groups of words. Sentiment analysis was also conducted to determine the impact and strength of negative or positive words that might change the context of the initial text analysis.

Part 3 – Student Engagement

As with the open-ended comments in Parts 1 and 2, students comments from both time periods were analysed together in order to establish recurring themes. The text modeling approach allows a quantitative value to be associated with each of the themes. Themes were also are correlated against the student's year, to establish if there were any age differences.

Reliability Analysis - Overall

As multiple Impact Coaches were applying the observation tool to various learning environments, it was important to establish the degree of consistency in how Impact Coaches were assessing the same teacher and student behaviors and interactions. Using Cohen's Kappa inter-rater reliability approach, findings showed that all three parts of the tool had high reliability coefficients: Part 1 (*kappa* = 0.67), Part 2 (*kappa* = 0.74) and Part 3 (*kappa* = 0.88). These results indicate that there was a strong degree of similarity across classroom observers, and therefore, there can be a high level of confidence and fidelity in the results presented in the *International Impact Report*.

Methods – Student Achievement

Sample

The sample comprised of Year 1 – 10 students (*n* = 2105) from 100 schools located in New Zealand and Australia. Data consisted of 67% of the students from Years Levels 1 - 6, and 33% from Years Levels 7 - 10. Just under half (44%) the students were male and 56% were female. The mean age of students was 10.60 (*SD* = 2.75) years.

Measures

It is hypothesized that the ultimate success of the Visible Learning^{plus} program is reflected through the improvements (relative and/or absolute) learning and achievement of students. The only requirement in order to accurately measure the programs impact is that a normed-referenced standardized test is adminstered that reflects the national curriculum.

The achievement data used in the *International Impact Report* comprised of the following five tests and subjects: Progressive Achievement Test (PAT) – Reading; Progressive Achievement Test (PAT) – Mathematics; Australian Victorian Essential Learning Standards (AusVELS); National Assessment Program – Literacy and Numeracy (NAPLAN) data.

Procedure

Tests were administered in early March (Time 1) and again in October/November (Time 2).

Data Analysis

In order to understand the impact of the Visible Learning^{plus} program it was necessary to compare the achievement data against the typical year of growth for a given target

population of students. To analyze the normative expectations for gain, an approach recommended by Kane (2004) was implemented. The analysis used the test scores, mean scale score and standard deviation, by Year Level from the national norming samples of each standardized tests. For each test, annual growth in achievement was measured by establishing the difference of mean scale score for both time periods for each Year Level. This was then converted into the difference to a standardized effect size by dividing it by the pooled standard deviation for each of time points pairing of Year Levels. This information was then aggregated across all the tests by taking the random-effect mean effect size for each Year Levels Time 1 and 2 gain using a weighting formula (see Hedges, 1982). Benchmarks were established so that there was a norm-based expectation of growth or change in the absence of the program[2]. These were then compared with the gains made by the students in the Visible Learning^{plus} program where there was a consistent pattern in both Reading and Mathematics. The average growth was found to be greatest in the earlier Year Levels (1 – 7), with gains declining amongst older students (Year Levels 8 - 10). However, this pattern has been found amongst each tests norm-referenced data, and has been observed in numerous randomized studies examining the impact of interventions (Shadish, Cook & Campbell, 2002).

Methods – Modeling Impact

Participants

The data for the across modelling across component analysis consisted of a sample of 10 schools from Central Australian that had participated in the Visible Learning^{plus} program over 2014. These schools had full data sets from the School Capability Assessment, Mindframes Survey, and students (n = 680) who had sat the PAT-R across both Time 1 and Time 2, and where there had been a minimum of 150 days between test administrations. There were 255 students across the secondary school levels (7-10) and 643 students at the primary/intermediate school levels (1-6). This sample consisted of 42% male and 58% female students.

Measures and Procedure

The Mindframes Survey, School Capability Assessments, and PAT-R measures were used in structural equation modelling (SEM). See above for detail relating to each of these measures.

[2] Due to low sample size of non-European students, the impacts of the program could not be compared with existing normed differences among subgroups.

Data Analysis

Structural Equation Modeling

A structural model was developed to establish the how components of the program impacted on student achievement. By modeling these relationships, it is possible to predict the effect that a latent variable can have on the outcome variable. For example, results in the *International Impact Report* showed that schools that had made significantly large gains in their school's Visible Learning capability, particularly in relation to establishing their vision and values, subsequently had made greater (relative) gains in their students' performance. These gains were particularly apparent for the low- and medium-performing students. Similarly, but to a slightly lesser degree, gains in the Mindframes Survey constructs were also related to a gain in student achievement.

Three rival models were developed to test if other characteristics were accounting for (or contributing to) the findings outlined in the report. It is always best practice to test whether another model might produce a better representation of the data. For example, one model tested if contextual effects such as the country or locality of the school (urban or rural) impacted on gains in student achievement results or results on the other latent variables. Another model was designed to test if student demographics, for example, ethnicity or school year level, had a direct or indirect effect on students' performance. The third model was structured to see if Mindframes Survey results were better interpreted nested within the School Capability Assessments.

Evaluation of Model Fit

On the basis of the recommendation of Hoyle and Panter (1995), this study included both absolute and incremental goodness-of-fit indexes for comparing models and analyzing invariance. The absolute fit index was represented by the chi-square statistic, although this statistic is problematic in terms of its power, especially with larger samples (see Marsh, Balla & McDonald, 1988). As Byrne (2001) noted, no matter how well a postulated a model is, it will always be falsely rejected given sufficient sample size. Thus, the chi-square statistic was established but not overly emphasized when interpreting the results. Also, because there was one nested models, the model fit for this was assessed using the chi-square difference test as well as other relative fit indexes. Following Conroy, Metzler, and Hofer (2003) suggestion, a greater emphasis was placed on the relative fit indexes, as these "are less sensitive to sample size and are more appropriate for evaluating badness of fit in regard to misspecification of factor loadings" (pg. 407). On the basis of extensive simulation studies, Cheung and Rensvold (2002) suggested that the criterion of -.01 change in the TLI between nested models was indicative that the hypothesis for change should not be rejected.

The incremental goodness-of-fit indexes use were the comparative fit index (CFI; Bentler, 1992), the TLI, and the root-mean-square error of approximation (RMSEA; Steiger & Lind, 1980). Both the CFI and the TLI have coefficient values ranging from 0 to 1.00, with values of .90 and higher traditionally viewed as representing good fit (Bentler, 1992). However, in a

revision of fit criteria, Hu and Bentler (1999) argued that this traditional value may fail to reject models that require respecifying, suggesting instead that .95 is a more accurate demonstration of good fit. Although there is conjecture around suggested fit values for the RMSEA, generally there is mediocre fit where values fall between .08 and .10 and reasonable fit where values are below .08 (Browne & Cudeck, 1993; Byrne, 2001; MacCallum, Browne, & Sugawara, 1996). Hu and Bentler suggested that an RMSEA less than or equal to .06 indicates good model fit.

Concluding Comments

The ongoing analysis of the information generated from the Visible Learning^{plus} program and its tools is vitally important for the ongoing validation and development of the Visible Learning^{plus} program. By understanding of the impact that the various components has provides empirical evidence of the impact that is being made at each major step of the program's delivery. In addition, it provides an insight into each school's capability, development and implementation of Visible Learning throughout the program.

As the datasets measuring the impact of Visible Learning^{plus} grows, so too will the ability to provide in-depth cross-sectional and comparative analysis within and across different international projects. In addition, as schools develop and evolve over the longer-term, their Visible Learning knowledge, understanding and practices will continue to evolve and become systematically embedded in their methods and processes. Such information will be invaluable in understanding the longitudinal impact of the program, particularly in relation to sustaining the impact of Visible Learning.

References

Adcock, E. P. & Phillips, G. W. (1997). Measuring school effects with hierarchical linear modeling: Data handling and modeling issues. *Multiple Linear Regression Viewpoints, 24*, 1-10.

Bentler, P. M. (1992). On the fit of models to covariances and methodology to the *Bulletin. Psychological Bulletin, 112,* 400–404.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Bryk, A. S. & Raudenbush, S. W. (1989). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. In Bock, R. D. (Ed). *Multilevel Analysis of Educational Data*. (pp.159-199). San Diego: Academic Press.

Byrne, B. M. (2001). *Structural equation modelling with Mplus: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 629-637.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272-299.

Hoyle, R., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts issues, and applications* (pp. 158–176). Thousand Oaks, CA: Sage.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Kirkpatrick, D. L. (1998). *Evaluating Training Programs* (2nd Edition), Donald L. Kirkpatrick, Berrett-Koehler Publishers, San Francisco, CA, 1998.

Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London, UK: Routledge.

Hattie, J. A. C. (2012). *Visible learning for teachers*. London, UK: Routledge.

Hu, L., & Bentler, P. M. (1999). Hierarchical confirmatory factor analysis of the revised Personal Style Inventory: Evidence for the multidimensionality problem for perfectionism. *Educational and Psychology Measurement, 61,* 421–432.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1,* 130–149.

Marsh, H. W., Balla, C. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103*, 391–410.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Steiger, J. H., & Lind, J. M. (1980, June). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Wothke, W. (2000). Longitudinal and multigroup modeling with missing data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), Modeling longitudinal and multiple-group data: Practical issues, applied approaches, and specific examples (pp. 219–240). Hillsdale, NJ: Erlbaum.